

5th Annual Data Miner Survey

– 2011 Survey Summary Report –

For more information contact
Karl Rexer, PhD
krexer@RexerAnalytics.com
www.RexerAnalytics.com

Outline

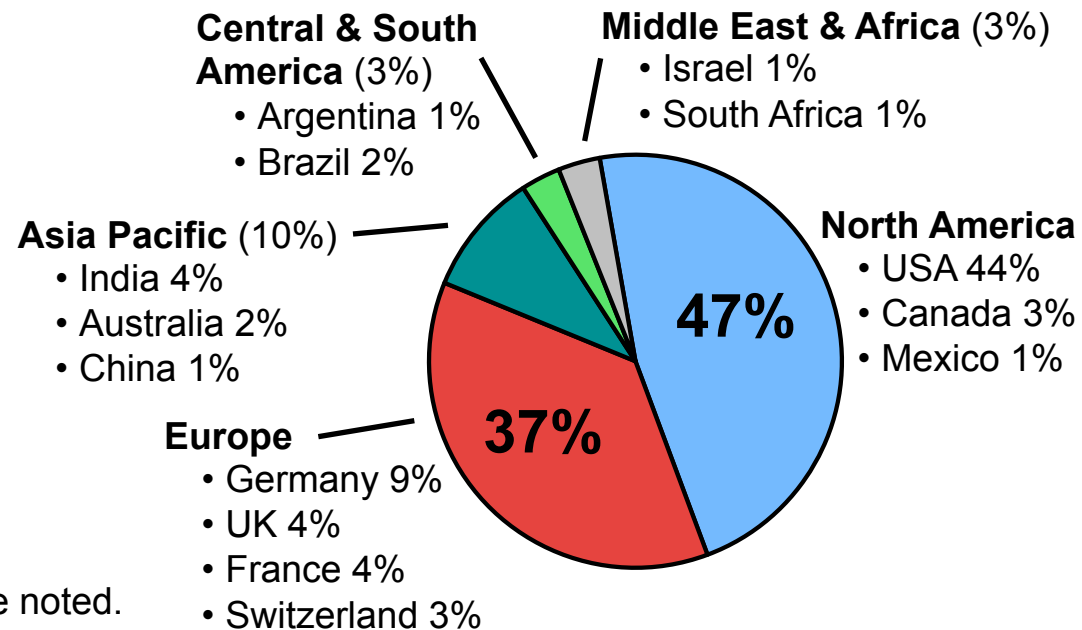
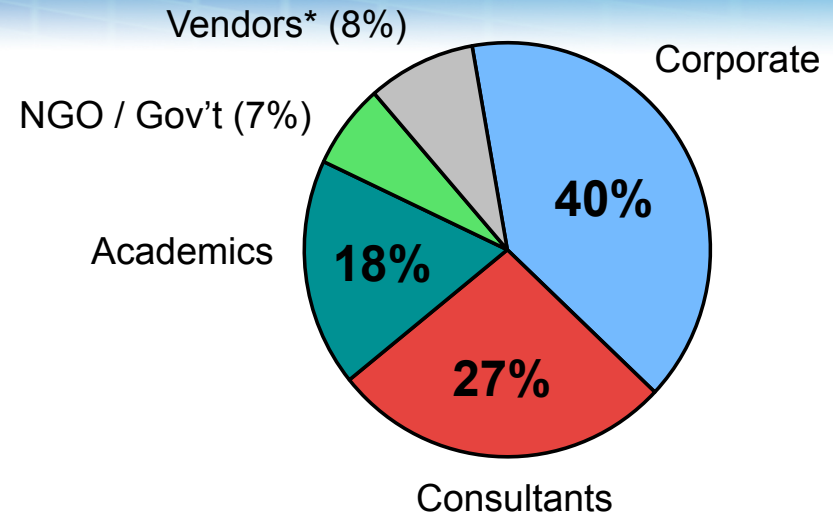
- Overview & Key Findings
- Where & How Data Miners Work
- Data Mining Tools: Usage & Satisfaction
- Goals, Challenges & Optimism about the Future
- Appendix: Rexer Analytics

Overview & Key Findings

2011 Data Miner Survey: Overview

Vendors are included in this analysis.

- 5th annual survey
- 52 questions
- 10,000+ invitations emailed, plus promoted by newsgroups, vendors, and bloggers
- Respondents: 1,319 data miners from over 60 countries
- Data collected in first half of 2011



*Data from software vendors is excluded from analyses in this presentation unless otherwise noted.

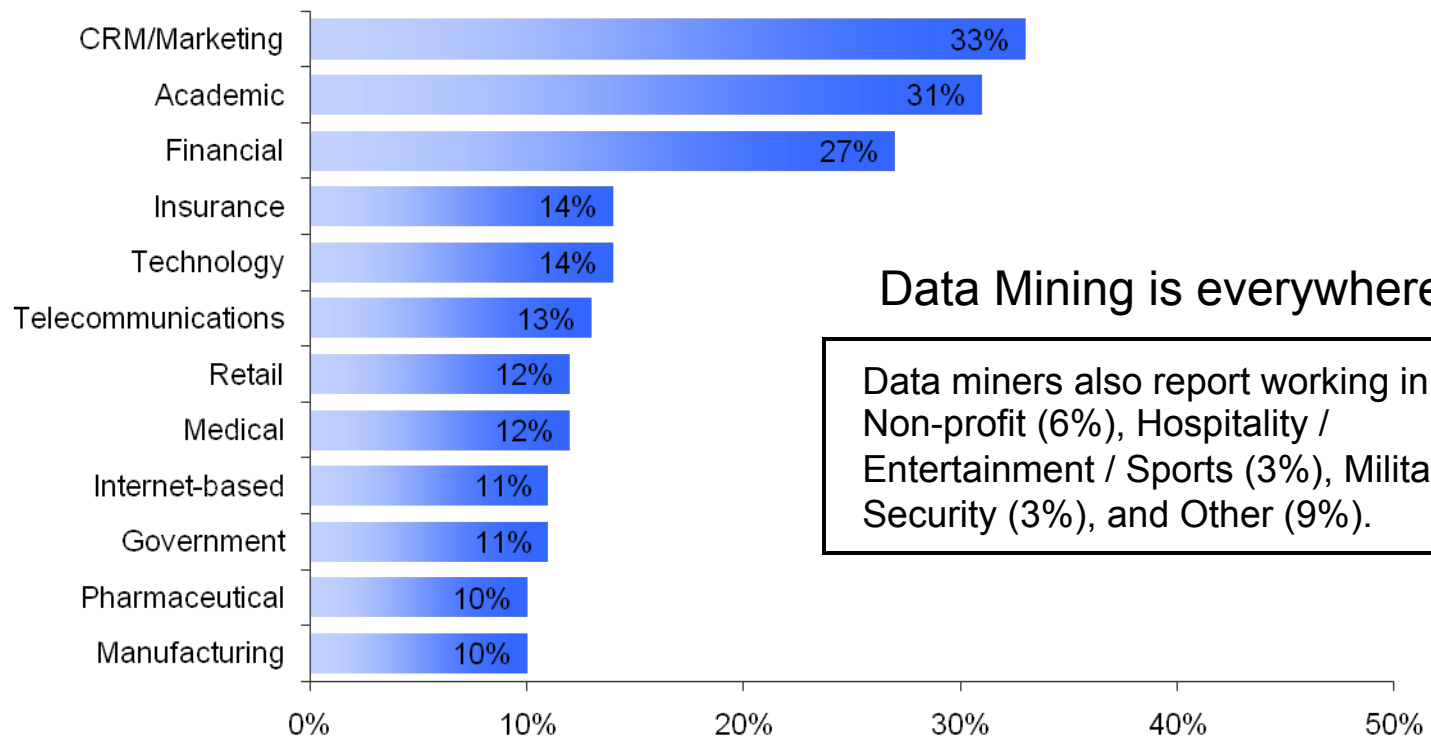
Key Findings

- **FIELDS & GOALS:** Data miners work in a diverse set of fields. CRM / Marketing has been the #1 field in each of the past five years. Fittingly, “improving the understanding of customers,” “retaining customers,” and other CRM goals continue to be the goals identified by the most data miners.
- **ALGORITHMS:** Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. However, a wide variety of algorithms are being used. A third of data miners currently use text mining and another third plan to in the future. Text mining is most often used to analyze customer surveys and blogs/social media.
- **TOOLS:** R continued its rise this year and is now being used by close to half of all data miners (47%). R users report preferring it for being free, open source, and having a wide variety of algorithms. Many people also cited R's flexibility and the strength of the user community. STATISTICA is selected as the primary data mining tool by the most data miners (17%). Data miners report using an average of 4 software tools overall. STATISTICA, KNIME, Rapid Miner, and Salford Systems received the strongest satisfaction ratings in 2011.
- **TECHNOLOGY:** Data Mining most often occurs on a desktop or laptop computer, and frequently the data is stored locally. Model scoring typically happens using the same software used to develop models.
- **VISUALIZATION:** Data miners frequently use data visualization techniques. More than four in five use them to explain results to others. MS Office is the most often used tool for data visualization. Extensive use of data visualization is less prevalent in the Asia-Pacific region than other parts of the world.
- **ANALYTIC CAPABILITY AND SUCCESS:** Only 12% of corporate respondents rate their company as having very high analytic sophistication. However, companies with better analytic capabilities are outperforming their peers. Respondents report analyzing analytic success via Return on Investment (ROI), and analyzing the predictive validity or accuracy of their models. Challenges to measuring analytic success include client or user cooperation and data availability/quality.

Where & How Data Miners Work

Data Miners are Working Everywhere

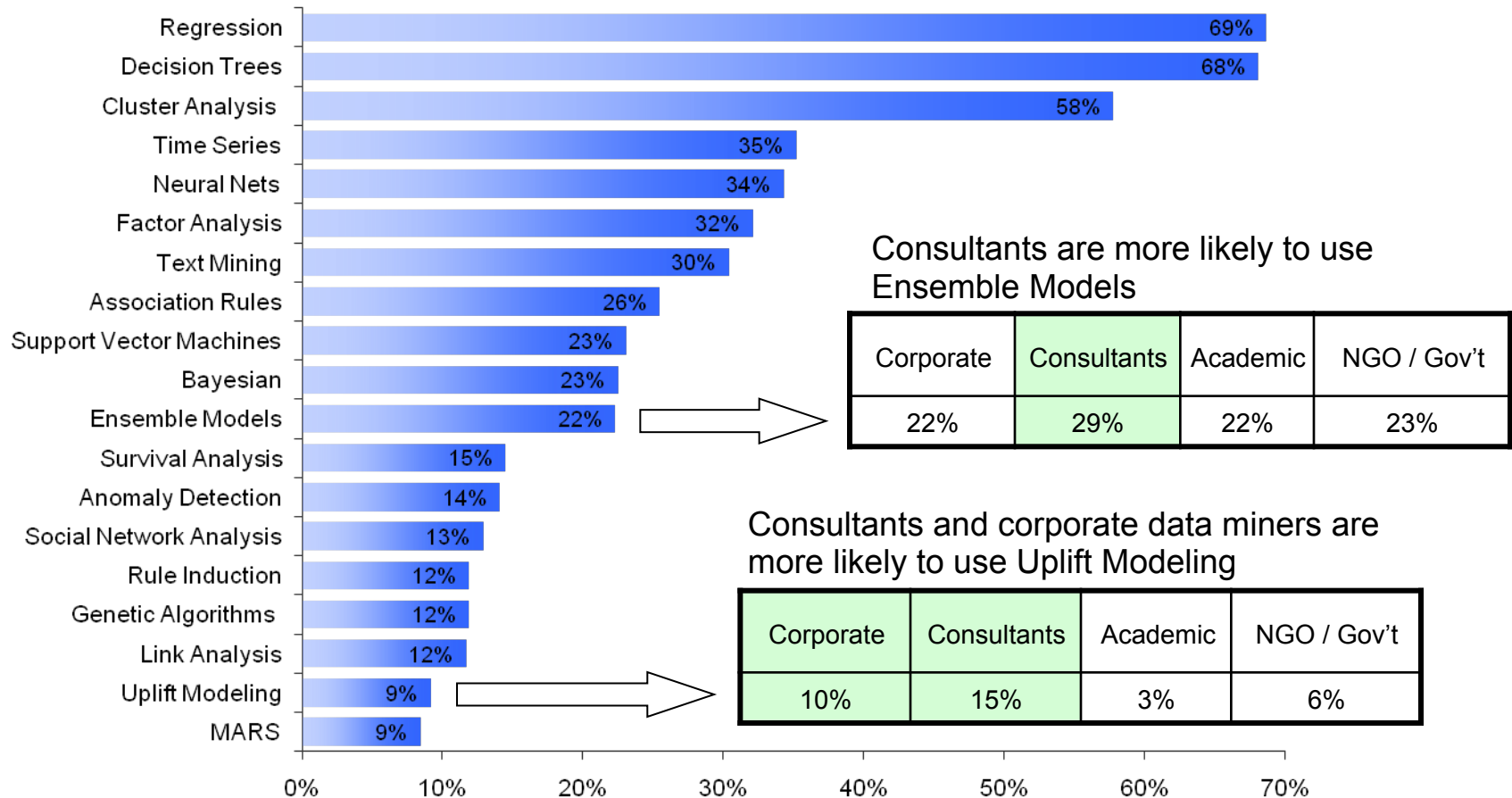
- More data miners report working in CRM / Marketing, Academia and Financial Services than any other fields.
 - These have been the three most commonly reported fields in each of the five annual Data Miner Surveys (2007-2011).
- Fewer data miners report working in CRM/Marketing this year (41% in 2010).
- Many data miners work in several fields.



Question: In what fields do you TYPICALLY apply data mining? (Select all that apply)

The Algorithms Data Miners are Using

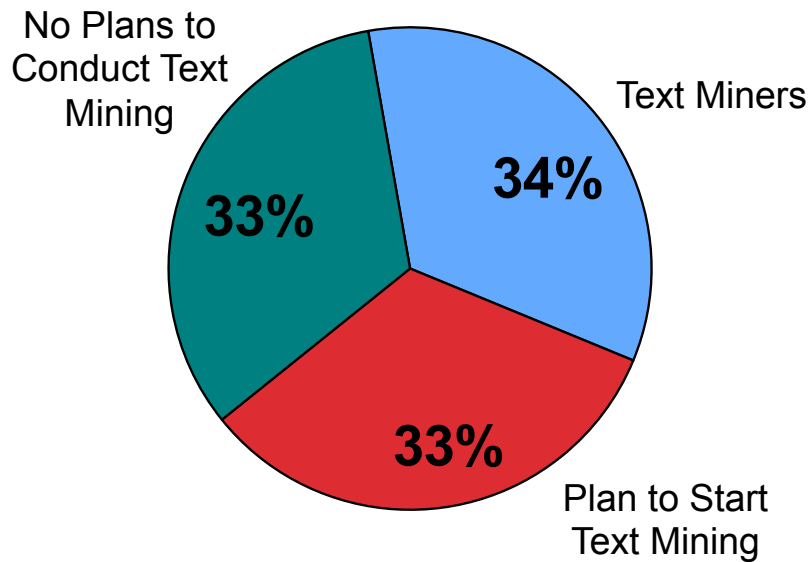
- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent over time.
- However, a wide variety of algorithms are being used.



Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

Text Mining

- About a third of data miners currently incorporate text mining into their analyses, while another third plan to do so.
- Academic data miners incorporate text mining into a larger proportion of projects.

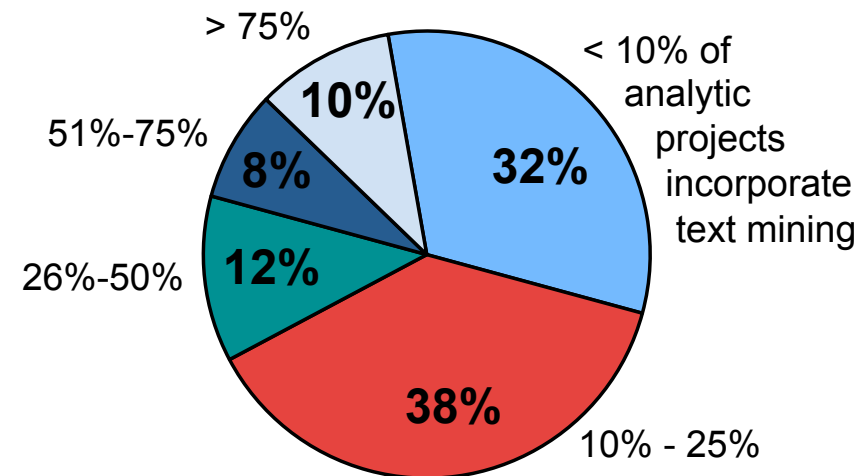


Question: Which is the best description of your use of text mining?

Text Material

Customer / market surveys	38%
Blogs and other social media	33%
E-mail or other correspondence	27%
News articles	25%
Scientific or technical literature	23%
Web-site feedback	22%
Online forums or review sites	21%
Contact center notes or transcripts	16%
Employee surveys	15%
Insurance claims or underwriting notes	15%
Medical records	11%
Point of service notes or transcripts	10%

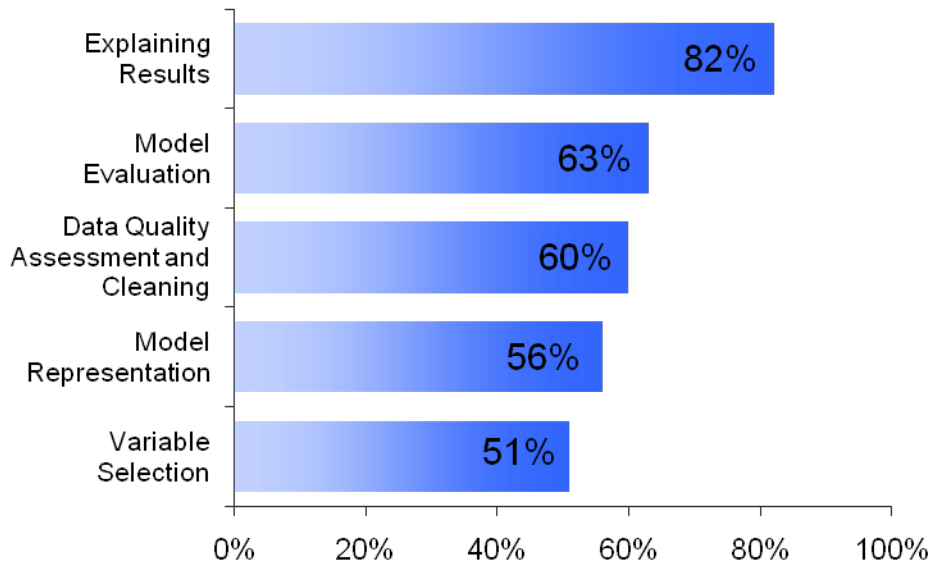
Question: In your text mining, what text material do you analyze or plan to analyze?



Question: What proportion of your analytic projects incorporate text mining?

Data Visualization

- Data miners frequently use data visualization techniques. More than four in five use them to explain results to others.
- MS Office is the most often used tool for data visualization.
- The extensive use of data visualization is less frequent in the Asia-Pacific region as compared to other parts of the world.



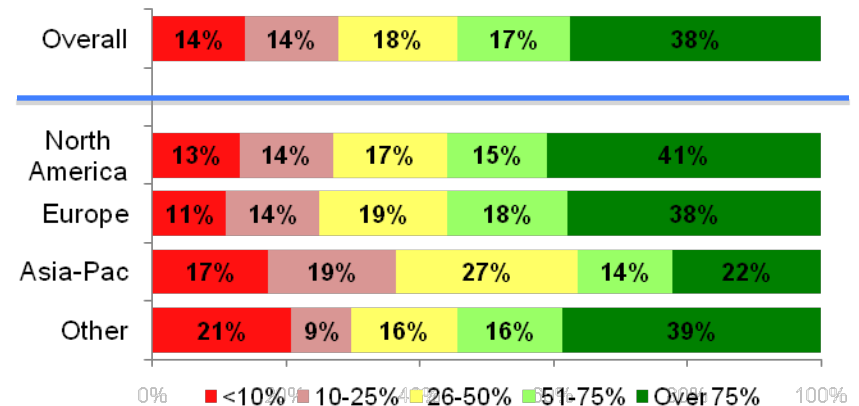
Question: In what areas do you employ graphical visualization during your analyses? (Check all that apply)

Primary Visualization Tools

MS Office	35%
R	28%
SAS	20%
STATISTICA	17%
IBM SPSS	15%
Rapid Miner	13%
Your own code	12%
IBM SPSS Modeler	11%
Matlab	10%
SAS Enterprise Miner	9%
Weka	8%
KNIME	8%

Question: Which one tool is your primary data visualization tool?

Proportion of Analytic Projects Incorporating Data Visualization

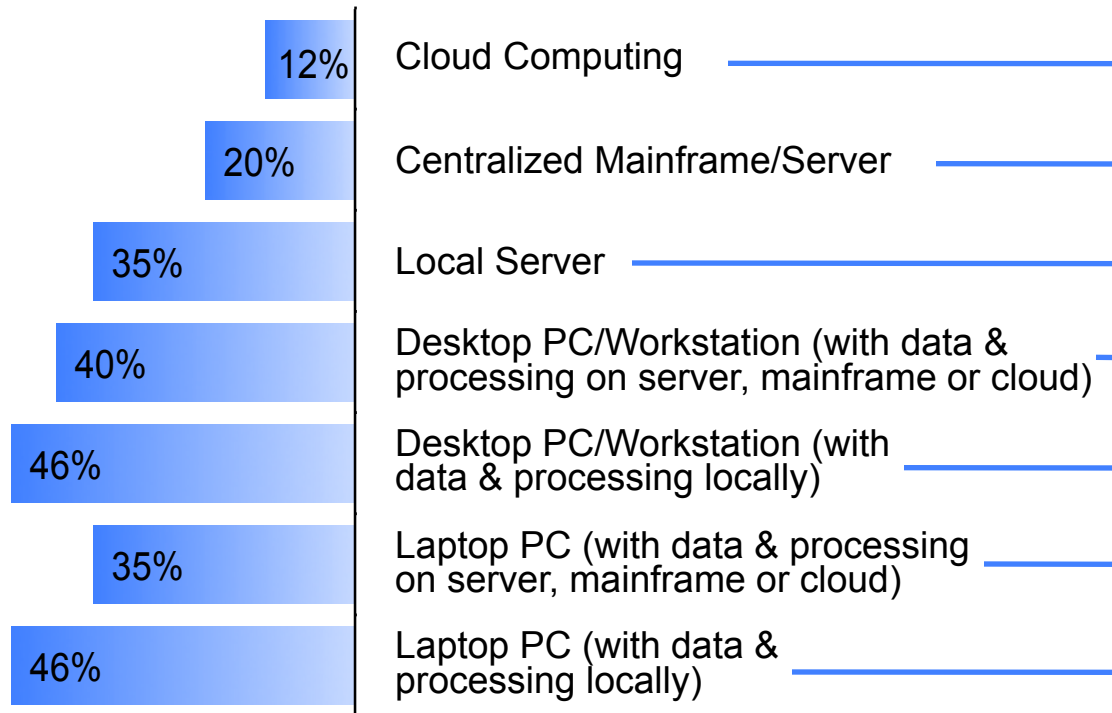


Question: What proportion of your analytic projects incorporate data visualization

Computing Environments

Vendors are included in this analysis.

- Most data mining happens on desktop and laptop computers.
- Frequently the data and processing is local (not on servers, mainframe or cloud).
- The proportion of data mining conducted on laptops has increased compared to 2010.



Overall

	Corporate	Consultant	Academic	NGO / Gov't	Vendor
Cloud Computing	9%	14%	10%	8%	25%
Centralized Mainframe/Server	20%	23%	12%	28%	24%
Local Server	32%	34%	33%	28%	50%
Desktop PC/Workstation (with data & processing on server, mainframe or cloud)	51%	36%	25%	45%	41%
Desktop PC/Workstation (with data & processing locally)	46%	43%	55%	57%	53%
Laptop PC (with data & processing on server, mainframe or cloud)	40%	36%	22%	24%	46%
Laptop PC (with data & processing locally)	35%	51%	59%	34%	56%

Question: What are the computing environments/platforms on which data mining/analytics occurs at your company/organization? (Check all that apply)

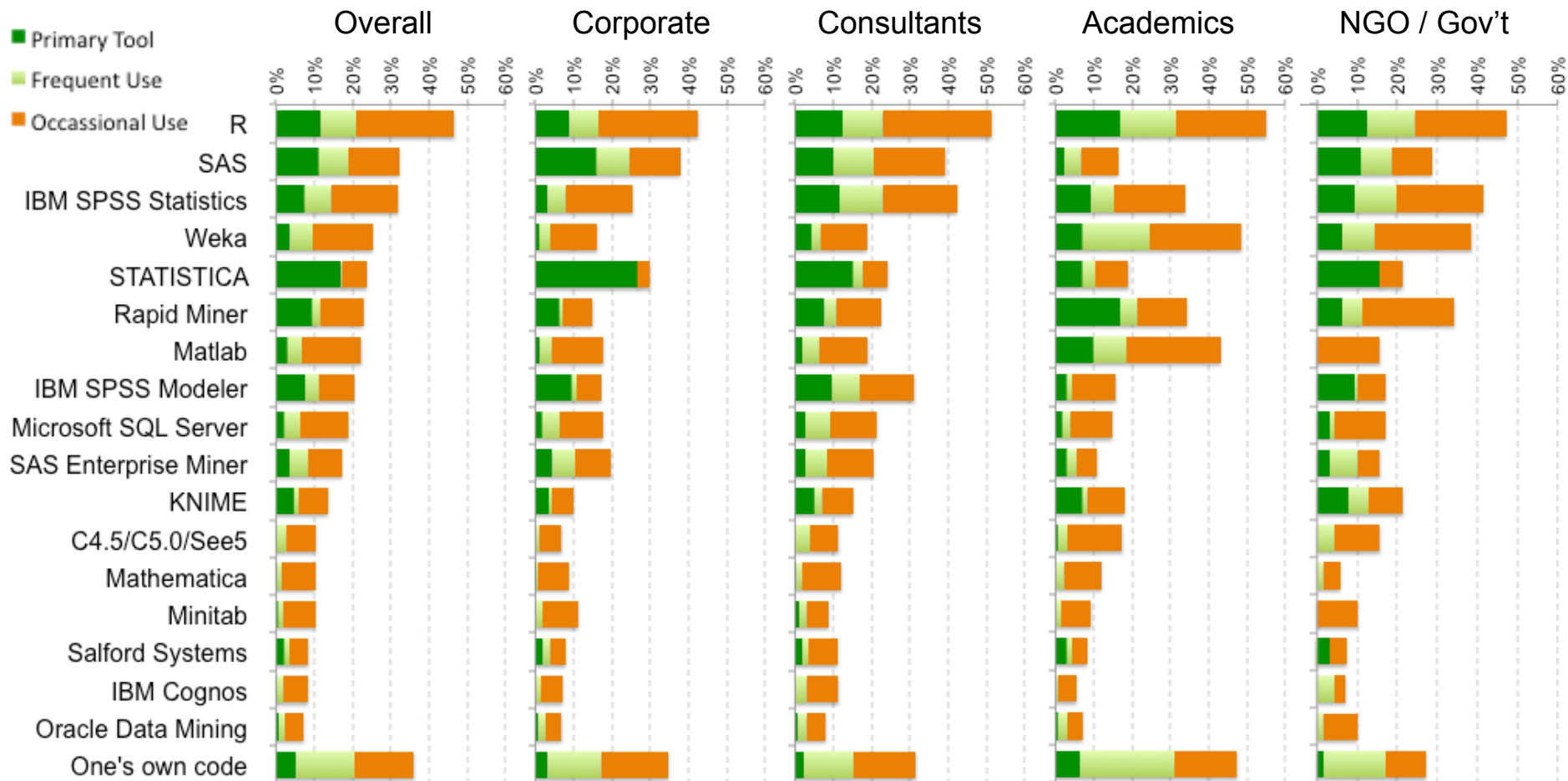
Data Mining Tools: Usage & Satisfaction

Data Mining Software

Survey Questions:

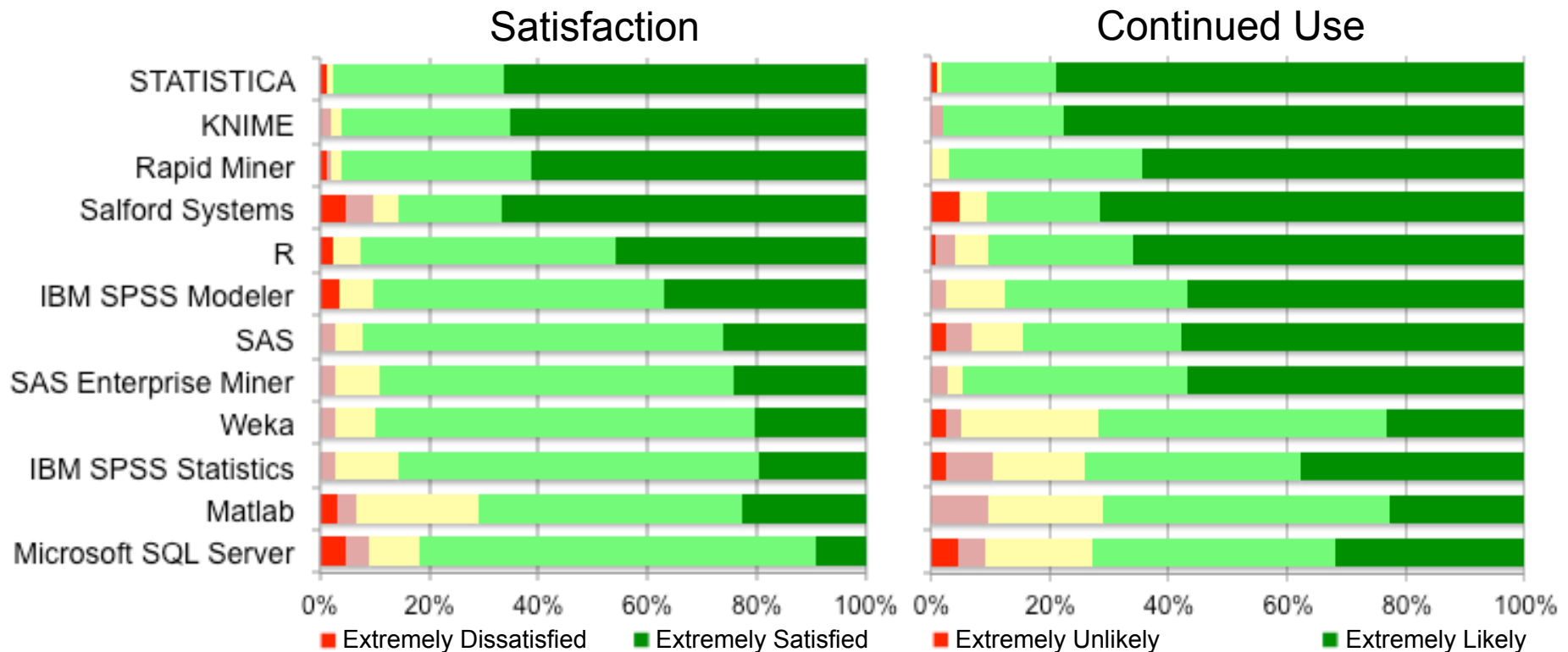
- What Data mining/analytic tools did you use in 2010? (rate each as “never”, “occasionally”, or “frequently”)
- What one Data Mining software package do you use most frequently?

- The average data miner reports using 4 software tools.
- R is used by the most data miners (47%).
- STATISTICA is the primary data mining tool chosen most often (17%).



Tools: Satisfaction & Continued Use

- STATISTICA, KNIME, Rapid Miner and Salford Systems received the highest satisfaction ratings.
- The users of these tools are also the most likely to continue using them as their primary tools for the next three years.



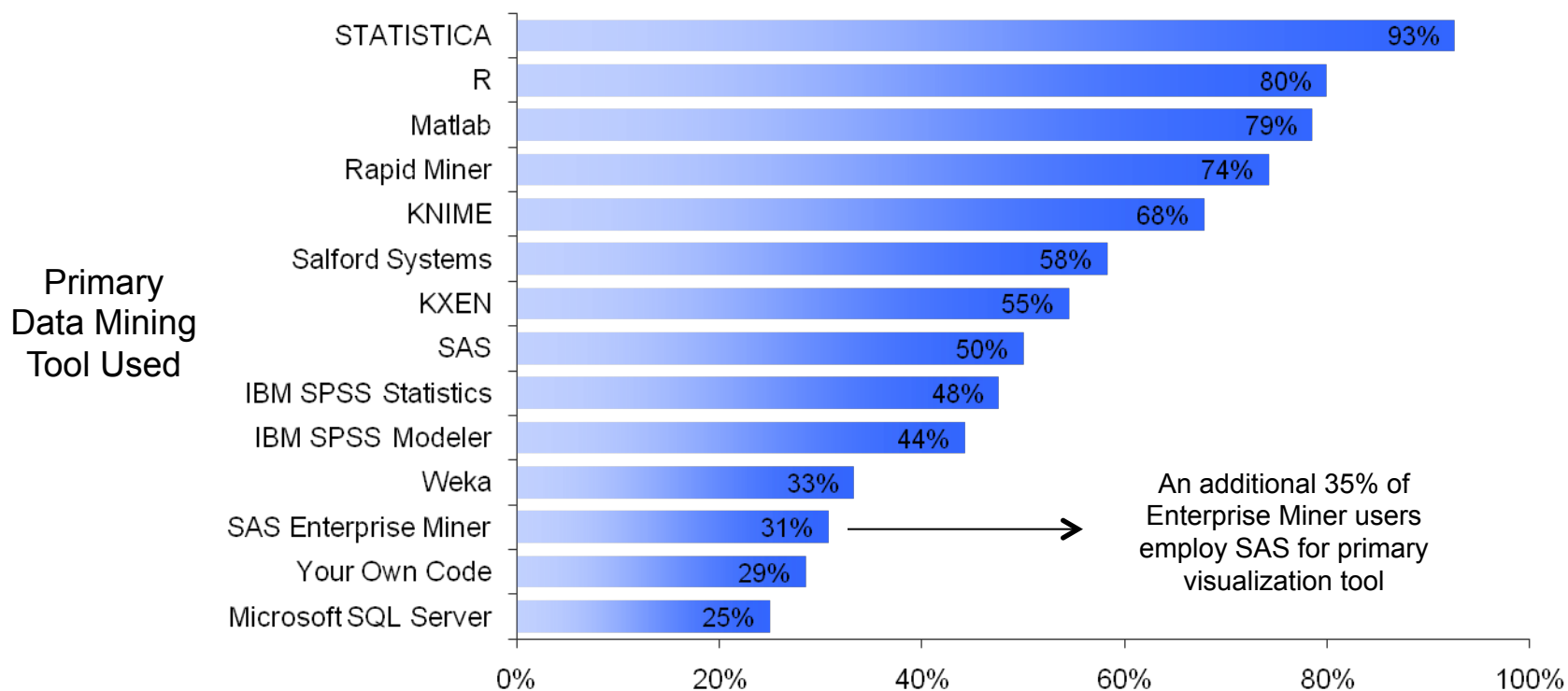
Satisfaction question: Please rate your overall satisfaction with your primary Data Mining software package.

Continued Use question: What is the likelihood that you will continue to use this tool as your primary Data Mining software package over the next 3 years?

Visualization Tools Used

- Over 9 in 10 of those who use STATISTICA as their primary data mining package also use it as their primary data visualization tool.
- R, Matlab, Rapid Miner, and KNIME also have a high percentage of users using the same tool for data visualization.

% Primary Data Mining Package Users that Identify the Same Package as Primary Visualization Tool

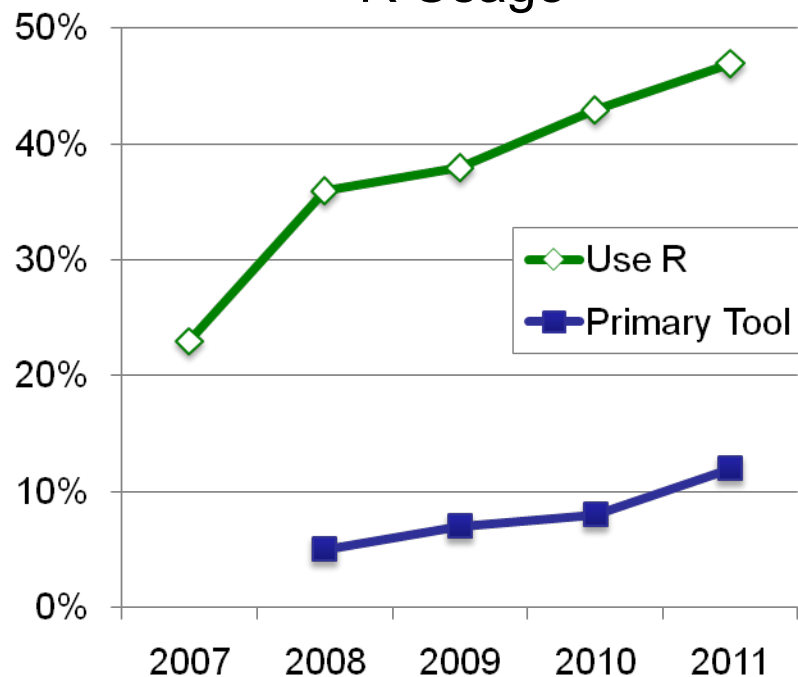


Question: What one data visualization tool do you use most frequently?

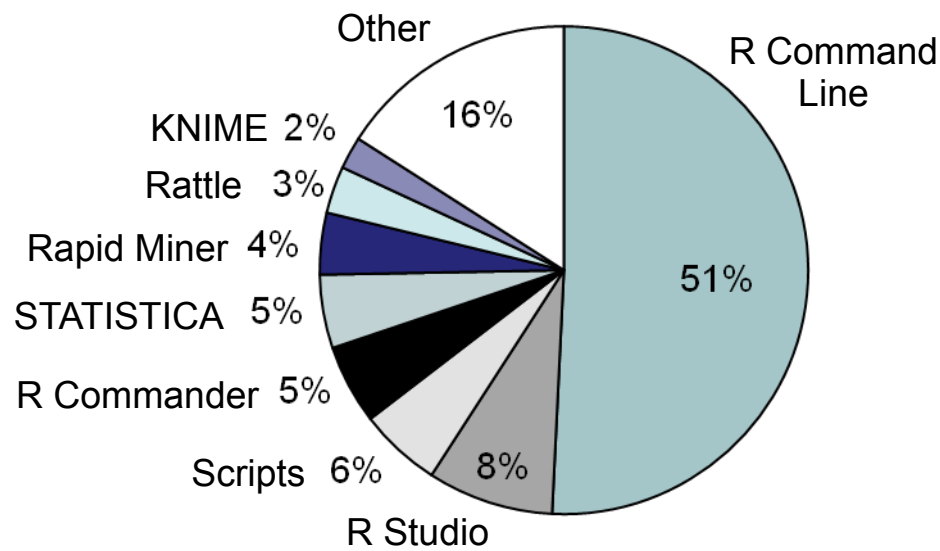
The Popularity of R Software is Growing Fast

- The proportion of data miners using R is rapidly growing.
 - R is also the #1 most used data mining tool (in both 2010 & 2011). Up from #5 in 2007.
- An increasing number of data miners consider R their primary tool.
 - R is now #2 in primary tool rankings. Up from #7 in 2008.
- Half of R users employ the command line interface. Among the rest, R Studio, scripts, R Commander, and STATISTICA are popular interfaces.

R Usage



R Interface



Question: If you use the R software package, what is your primary interface to R?

Insights from R Users

- 225 R users shared information about how and why they are using R. They provided an enormous wealth of useful and detailed information. We strongly encourage anyone with an interest in R to read the complete verbatim list of these R user's comments on the Rexer Analytics website: www.RexerAnalytics.com/DMSurvey2011_R-Comments.
- Here are a few examples of their comments. Many of the comments are much longer and more detailed.

Why data miners use R (Pros)

The reasons expressed by the most people focused on R being free, open source, and having a wide variety of algorithms. Many people also cited R's flexibility and the strength of the user community.

- "Best variety of algorithms available, biggest mindshare in online data mining community, free/open source."
- "Excellent graphics, wide variety of routines available, runs on multiple platforms (including Linux), many graphical interfaces available (some better than others for specific purpose), flexibility of programming language and interface to various databases."

Question: If you use R, please tell us more about your use of R. For example, tell us why you have chosen to use R, why you use the R interface you identified in the previous question, the pros and cons of R, or tell us how you use R in conjunction with other tools. (text box provided for response)

Insights from R Users (continued)

Cons of using R

A number of people mentioned a steep learning curve, frustrations with the interface, slow performance, memory limitations, and lack of support.

- "The main drawback to R, in my opinion, is that R loads in live memory all the work space it is linked to which is a big waste of time and memory and makes it difficult to use R in a multi-users environment where typical projects consist of several very large data sets."
- "Compared to some latest commercial software I've evaluated, R is sluggish for certain tasks, and can't handle very large datasets (mainly because I do not have a 64-bit machine to work with). On top of that, to be really productive with R, one needs to learn other languages, e.g., SQL, but that's just how things are."

Why R users select their chosen R interface

- "I mostly work with the command line, but I am moving towards RStudio because it's available both as a desktop application and a browser-based client-server tool set. I occasionally use Rcmdr."
- "I find the R GUI the most flexible way to use it. On occasion I've used JGR and Deducer, but I've generally found it more convenient to use the GUI."

How people use R in conjunction with other tools

- "I use R in conjunction with Matlab mostly, programming my personalized algorithms in Matlab and using R for running statistical test, ROC curves, and other simple statistical models."
- "RapidMiner offers access to R. The advantage of R is that a new algorithm can easily be developed -- and then be applied within RapidMiner."

Satisfaction with Tools: Details

- STATISTICA received strong ratings across all dimensions.

	Overall	IBM SPSS Statistics	IBM SPSS Modeler	KNIME	R	Rapid Miner	SAS	SAS Enterprise Miner	STATISTICA (StatSoft)	Weka
Quality and accuracy of model performance	4.34	3.93	4.32	4.32	4.43	4.37	4.34	4.26	4.69	4.21
Dependability/Stability of software	4.25	4.19	4.05	4.43	4.34	4.22	4.32	4.44	4.56	3.73
Variety of available algorithms	4.20	3.69	4.30	4.48	4.72	4.54	4.01	4.00	4.63	4.33
Ease of use	4.19	4.28	4.60	4.76	3.58	4.47	3.69	4.00	4.49	4.06
Ability to automate repetitive tasks	4.17	3.75	3.96	4.39	4.39	4.40	4.31	4.00	4.45	3.71
Data manipulation capabilities	4.15	4.00	4.32	4.53	4.10	4.27	4.45	3.82	4.41	3.52
Quality of output / Ease of interpretation	4.10	3.91	4.04	4.39	4.04	4.36	3.69	4.00	4.53	3.66
Good metrics of model quality	4.10	3.85	3.96	4.05	4.13	4.28	4.08	4.18	4.50	3.85
Good variable discovery, profiling and selection	4.03	3.70	4.06	4.17	3.98	4.33	3.81	4.35	4.44	3.69
Quality of user interface	4.03	4.11	4.53	4.62	3.36	4.45	3.58	3.91	4.49	3.59
Ease of model deployment (scoring models to other data sets)	4.03	3.61	4.13	4.43	3.82	4.20	3.90	4.21	4.46	3.77
Speed	4.02	3.84	4.13	4.12	3.58	3.90	4.08	3.97	4.48	3.53
Data quality assessment and data preparation capabilities	4.00	3.94	4.22	4.33	3.76	4.20	4.05	3.68	4.38	3.53
Ability to handle very large data sets	3.99	3.82	4.21	4.35	2.95	3.74	4.41	4.44	4.58	3.03
Ability to modify algorithm options to fine-tune analyses	3.95	3.17	3.59	3.98	4.33	4.23	3.97	3.97	4.33	3.88
Enables mining within one's database	3.94	3.54	4.26	4.12	3.75	4.10	3.92	4.00	4.19	3.61
Ability to easily incorporate data at different levels of granularity (e.g. transaction data and customer data)	3.90	3.56	4.06	4.24	3.77	3.99	4.14	3.94	4.24	3.29
Useful help menu, demos and tutorials	3.87	3.83	3.99	3.93	3.68	3.90	3.76	3.79	4.35	3.61
Strong graphical visualization of models	3.83	3.24	3.68	3.88	4.14	4.28	3.02	3.88	4.62	3.28
Cost of software	3.79	3.16	3.00	4.93	4.90	4.82	2.33	2.74	3.90	4.88

Mean satisfaction rating on 1-5 scale

 Higher Satisfaction

 Lower Satisfaction

Question: Rate how satisfied you are with the performance of your primary data mining package (identified earlier) on each of these factors.

Factors Most Related to Primary Tool Satisfaction

- The simple correlations between detailed satisfaction items and overall primary tool satisfaction reveal the factors most closely related to primary tool satisfaction.

	Correlation with Overall Satisfaction									
	Overall		Corporate		Consultant		Academic		NGO / Gov't	
	Correlation	Rank	Correlation	Rank	Correlation	Rank	Correlation	Rank	Correlation	Rank
Good variable discovery, profiling and selection	.419	1	.453	1	.401	5	.471	6	.306	5
Quality of output / Ease of interpretation	.399	2	.426	4	.419	2	.390	12	.241	10
Good metrics of model quality	.395	3	.374	8	.362	10	.535	1	.326	2
Strong graphical visualization of models	.392	4	.391	6	.414	3	.478	3	.278	8
Quality and accuracy of model performance	.380	5	.372	10	.389	8	.478	4	.184	16
Ability to modify algorithm options to fine-tune analyses	.377	6	.436	2	.309	18	.353	14	.455	1
Data quality assessment and data preparation capabilities	.376	7	.427	3	.322	17	.438	8	.234	12
Variety of available algorithms	.375	8	.363	11	.362	11	.500	2	.311	4
Dependability/Stability of software	.365	9	.343	13	.469	1	.375	13	.300	6
Ability to easily incorporate data at different levels of granularity (e.g. transaction data and customer data)	.356	10	.346	12	.393	7	.440	7	.206	14
Ease of model deployment (scoring models to other data sets)	.353	11	.373	9	.329	14	.408	10	.235	11
Quality of user interface	.352	12	.340	14	.398	6	.475	5	.157	17
Data manipulation capabilities	.349	13	.333	15	.408	4	.391	11	.248	9
Ease of use	.340	14	.392	5	.323	15	.302	16	.285	7
Ability to automate repetitive tasks	.328	15	.313	16	.347	12	.411	9	.317	3
Speed	.323	16	.381	7	.389	9	.252	19	.123	18
Ability to handle very large data sets	.274	17	.305	17	.323	16	.268	18	.104	20
Enables mining within one's database	.272	18	.234	20	.346	13	.303	15	.118	19
Useful help menu, demos and tutorials	.262	19	.267	18	.283	19	.270	17	.233	13

Data Mining Tools: Strengths & Weaknesses

- Tool strengths and weaknesses were identified by the satisfaction ratings of data miners who considered each tool to be their primary data mining tool.

	IBM SPSS Statistics	IBM SPSS Modeler	Knime	R	Rapid Miner	SAS	SAS Enterprise Miner	STATISTICA (StatSoft)	Weka
Top 3 Strengths	<ol style="list-style-type: none"> 1) Ease of use 2) Dependability/ Stability of software 3) Quality of user interface 	<ol style="list-style-type: none"> 1) Ease of use 2) Quality of user interface 3) Data manipulation capabilities 	<ol style="list-style-type: none"> 1) Ease of use 2) Quality of user interface 3) Data manipulation capabilities 	<ol style="list-style-type: none"> 1) Variety of available algorithms 2) Quality and accuracy of model performance 3) Ability to automate repetitive tasks 	<ol style="list-style-type: none"> 1) Variety of available algorithms 2) Ease of use 3) Quality of user interface 	<ol style="list-style-type: none"> 1) Data Manipulation capabilities 2) Ability to handle very large data sets 3) Quality and accuracy of model performance 	<ol style="list-style-type: none"> 1) Ability to handle very large data sets 2) Dependability / Stability of software 3) Good variable discovery, profiling and selection 	<ol style="list-style-type: none"> 1) Quality and accuracy of model performance 2) Variety of available algorithms 3) Strong graphical visualization of models 	<ol style="list-style-type: none"> 1) Variety of available algorithms 2) Quality and accuracy of model performance 3) Ease of use
Top 3 Weaknesses	<ol style="list-style-type: none"> 1) Ability to modify algorithm options to fine-tune analyses 2) Strong graphical visualization of models 3) Enables mining within one's database 	<ol style="list-style-type: none"> 1) Ability to modify algorithm options to fine-tune analyses 2) Strong graphical visualization of models 3) Ability to automate repetitive tasks 	<ol style="list-style-type: none"> 1) Strong graphical visualization of models 2) Useful help menu, demos and tutorials 3) Ability to modify algorithm options to fine-tune analyses 	<ol style="list-style-type: none"> 1) Ability to handle very large data sets 2) Quality of user interface 3) Speed 	<ol style="list-style-type: none"> 1) Ability to handle very large data sets 2) Speed 3) Ability to easily incorporate data at different levels of granularity 	<ol style="list-style-type: none"> 1) Strong graphical visualization of models 2) Quality of user interface 3) Quality of output/ease of interpretation 	<ol style="list-style-type: none"> 1) Data quality assessment and data preparation capabilities 2) Useful help menu, demos and tutorials 3) Data Manipulation capabilities 	<ol style="list-style-type: none"> 1) Enables mining within one's database 2) Ability to easily incorporate data at different levels of granularity 3) Ability to modify algorithm options to fine-tune analyses 	<ol style="list-style-type: none"> 1) Ability to handle very large data sets 2) Strong graphical visualization of models 3) Ability to easily incorporate data at different levels of granularity

Note: Strengths and weaknesses determined by mean on 5-point satisfaction scale.

Question: Rate how satisfied you are with the performance of your primary data mining package (identified earlier) on each of these factors.

Goals, Challenges & Optimism about the Future

Goals for Analyses

- The goals for data mining analyses are diverse.
- Several CRM goals are high on the list. More than a third of data miners indicate that they are using data mining to improve the understanding of customers.
- Some data mining tools are used for a wide range of goals, and others have more specific, niche uses.

	Overall	IBM SPSS Statistics	IBM SPSS Modeler	KNIME	R	Rapid Miner	SAS	SAS Enterprise Miner	STATISTICA (StatSoft)	Weka
Total Number of Goals	3.7	3.8	5.6	3.2	3.4	3.2	4.5	5.2	3.1	2.5
Improving understanding of customers	33%	45%	54%	20%	32%	25%	50%	54%	19%	23%
Retaining customers	30%	28%	59%	16%	24%	14%	44%	43%	25%	15%
Market research / survey analysis	29%	54%	40%	31%	28%	20%	40%	38%	21%	18%
Scientific discovery/ advancement	27%	18%	11%	55%	44%	35%	13%	14%	18%	51%
Selling products / services to existing customers	23%	21%	48%	16%	15%	21%	31%	43%	12%	13%
Acquiring customers	23%	26%	46%	14%	16%	10%	38%	35%	17%	13%
Improving direct marketing programs	22%	28%	48%	14%	15%	13%	38%	35%	11%	5%
Improving customer experiences	22%	26%	37%	20%	18%	18%	35%	30%	14%	21%
Risk management / credit scoring	22%	21%	32%	18%	15%	17%	32%	35%	27%	5%
Fraud detection or prevention	21%	6%	38%	16%	15%	16%	23%	41%	32%	10%
Sales forecasting	19%	22%	30%	10%	16%	16%	27%	32%	18%	3%
Price optimization	14%	14%	16%	6%	15%	11%	17%	27%	15%	3%
Medical advancement / drug discovery / biotech / genomics	12%	10%	12%	24%	19%	7%	6%	5%	9%	8%
Investment planning / optimization	11%	9%	10%	2%	11%	12%	7%	8%	21%	8%
Manufacturing improvement	10%	3%	14%	2%	7%	14%	5%	5%	19%	3%
Website or search optimization	8%	5%	9%	8%	7%	11%	10%	8%	2%	5%
Supply chain optimization	7%	4%	12%	2%	8%	7%	3%	8%	8%	3%
Software optimization	7%	3%	4%	6%	10%	12%	5%	11%	3%	10%
Collections	6%	8%	6%	4%	2%	3%	14%	19%	3%	0%
Human resource applications	4%	13%	6%	4%	2%	1%	3%	3%	6%	0%
Information security	4%	5%	6%	8%	4%	2%	2%	5%	4%	5%
Language understanding	4%	3%	4%	8%	3%	12%	2%	5%	0%	8%
Criminal or terrorist detection	3%	1%	11%	2%	3%	3%	0%	3%	1%	15%
Natural resource planning or discovery	3%	5%	2%	4%	7%	7%	0%	3%	2%	0%
Fundraising	3%	5%	6%	2%	0%	2%	2%	14%	1%	5%
Reducing email spam	2%	1%	1%	0%	2%	3%	3%	0%	1%	3%

Question: What were the goals of your analyses in 2010? (Select all that apply)

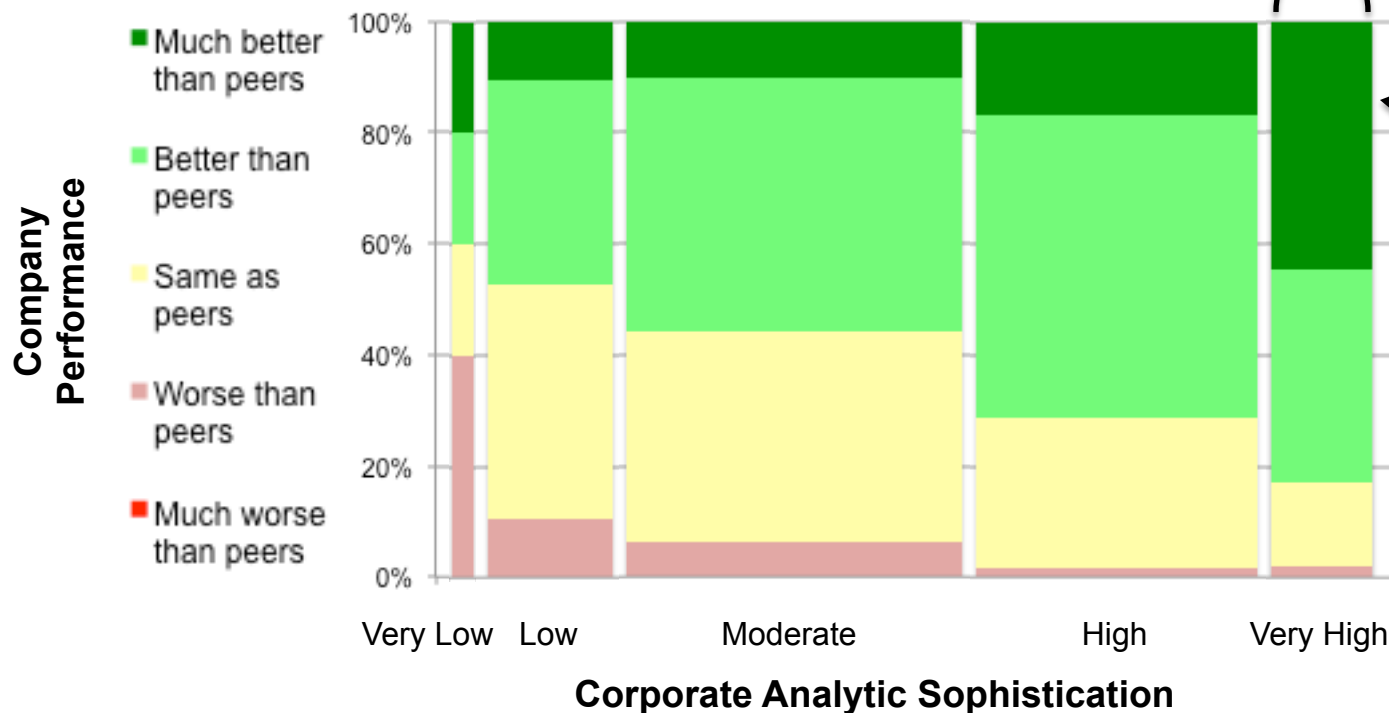
Higher Frequency

Lower Frequency

Room for Improvement ... And it Matters!

- Analytic capability:
 - There's room to improve if we are going to "Compete on Analytics".
- Analytic capabilities boost company performance.

Only 12% of corporate respondents rate their company as having very high analytic sophistication.



Companies with better analytic capabilities are outperforming their peers!

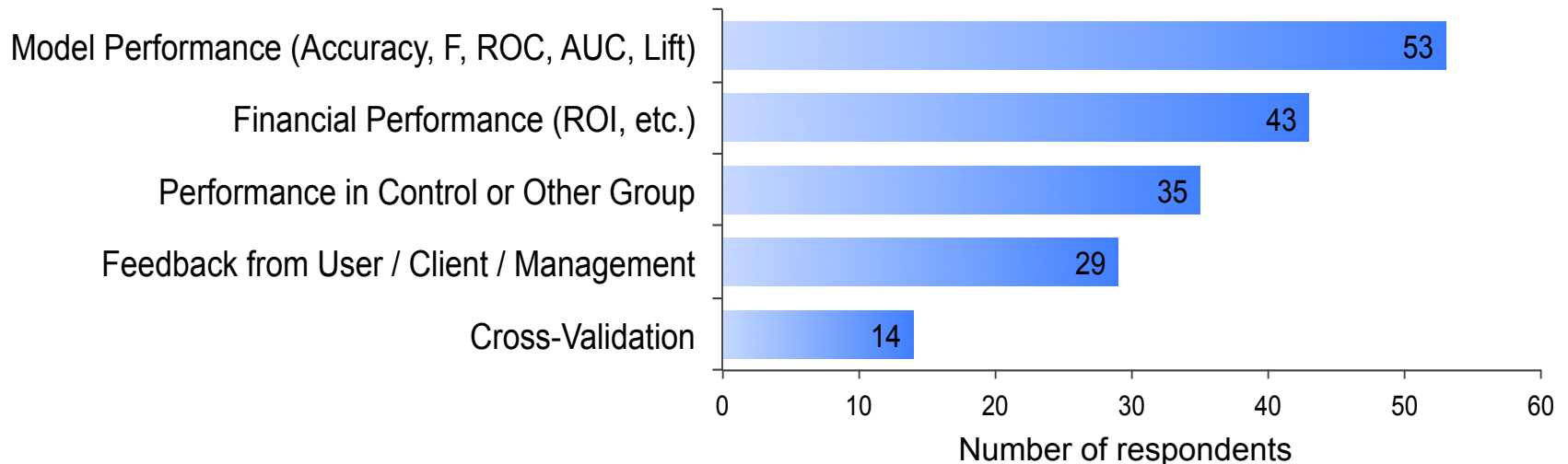
Caution: this is self report data & correlation analysis.

Analytic Capability Question: In general, with what degree of sophistication does your company / organization approach analytic problems?

Company Performance Question: Which statement best describes the recent performance of your company / organization?

Measuring Analytic Success

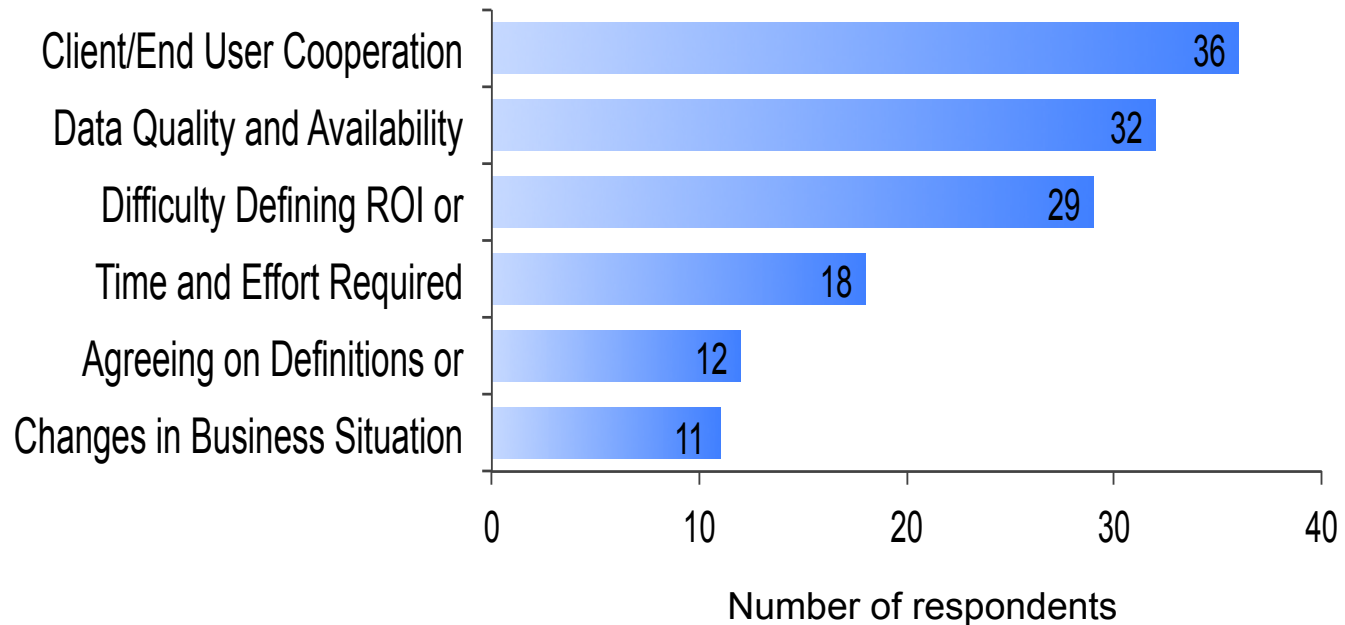
- Survey respondents shared their best practices in ways to measure analytic success (an open-ended survey question).
- Model performance (accuracy, F, ROC, AUC, lift) and financial performance (ROI and other financial measures) were the best practice methods described by the most data miners.
- Many data miners use multiple methods, and a wide range of methods are being used (131 data miners described methods that fall outside the categories graphed below).
- For a complete list of respondents' ideas on best practices in measuring analytic success, see www.rexeranalytics.com/DMSurvey2011_MeasuringSuccess.



Question: Please share your best practices concerning how you measure analytic project performance / success. (text box provided for response)

Challenges to Measuring Analytic Success

- Survey respondents shared their ideas about challenges in measuring analytic success (an open-ended survey question).
- Client or user cooperation, and data availability/quality were cited most frequently.



Question: Please describe the main challenges you've experienced in measuring analytic project performance/success. If you've overcome the challenge, please describe how you accomplished this. (text box provided for response)

There's Strong Demand

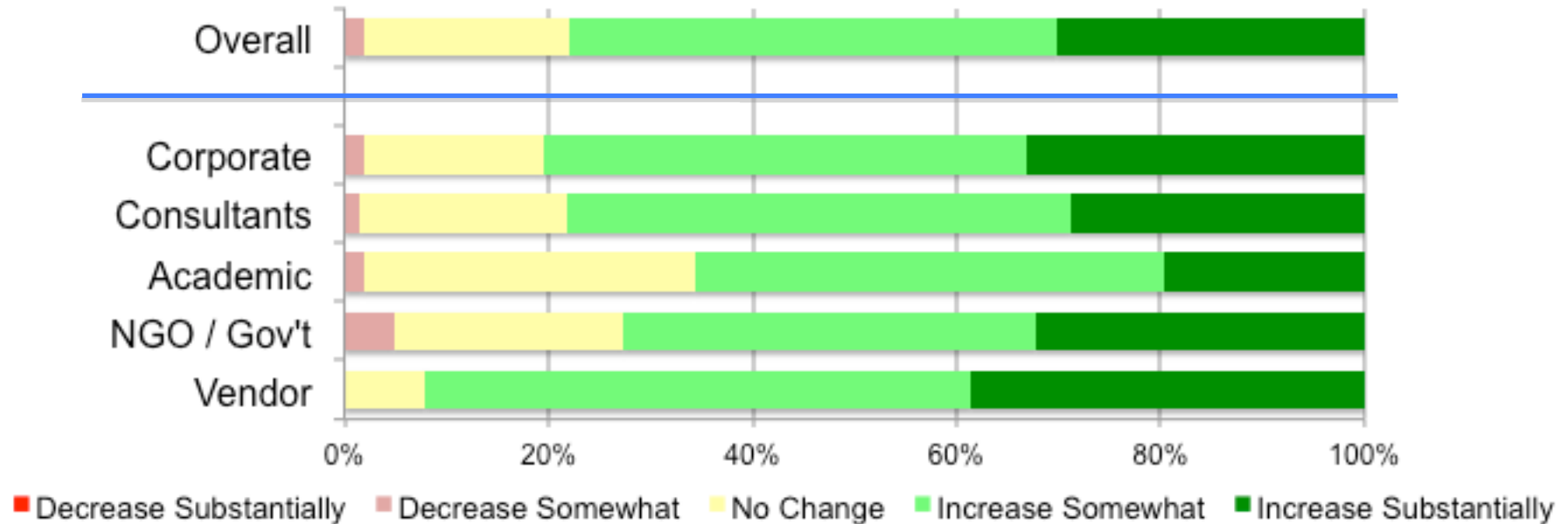
Vendors are included in this analysis.

Data miner hiring is very strong*.

Company use of data mining is increasing.

- 78% of data miners foresee increases in the number of data mining projects.
- This is consistent with similar increases projected last year.
- Data miners working in diverse settings share this optimism.

Number of Data Mining Projects Projected in 2011

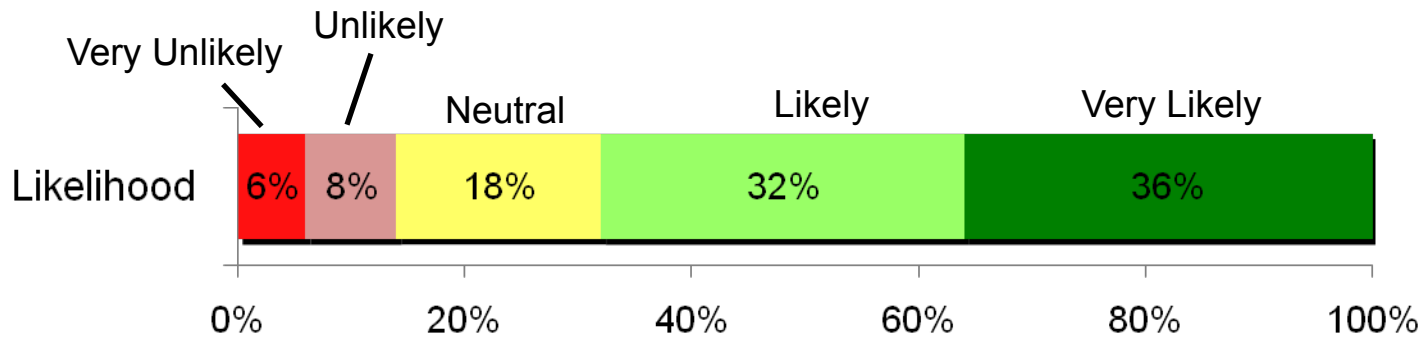
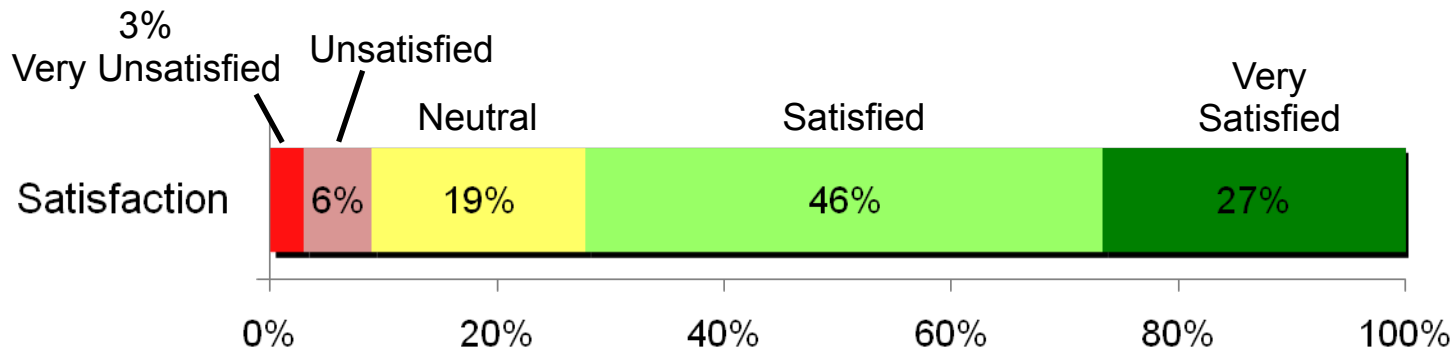


Question: How will the number of data mining projects your organization conducts in 2011 compare to what has been typical in the past few years?

* Multiple sources: Use of "data mining" in online job ads, KDnuggets job listings, recruiters, salary reports.

We Enjoy Our Work

- Data miners are generally satisfied with their jobs, with more than a quarter reporting being “very satisfied.”
- 68% report being likely to remain with their current employer for the next two years.



Questions: What is your current level of job satisfaction?
How likely are you to remain with your current employer for the next two years?

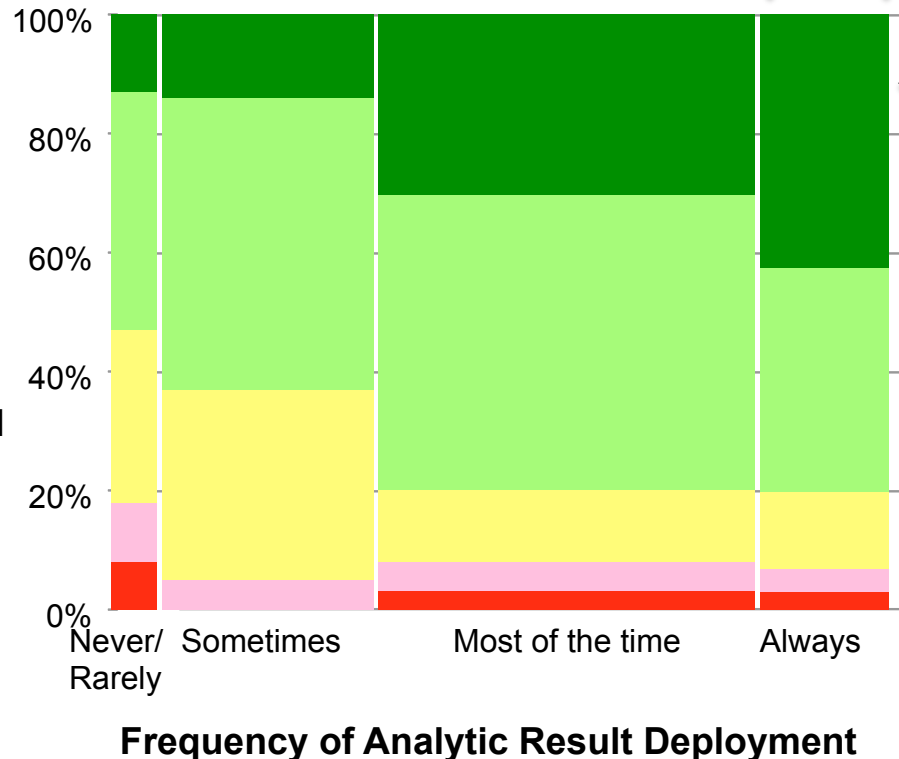
Utilization of Analytics Influences Our Job Satisfaction

- At about 1 in 3 organizations, deployment of analytic results is not a common occurrence.
- Data miners have higher job satisfaction at organizations where results are commonly deployed.

Only 17% of respondents rate their company as “always” deploying/utilizing analytic results.

Job Satisfaction

- Very Satisfied
- Satisfied
- Neutral
- Unsatisfied
- Very Unsatisfied



Companies which deploy analytic results more frequently produce more satisfied analysts.

Question: What is your current level of job satisfaction?

Question: How often are results of your analytics deployed and/or utilized?

Reasons for Non-Deployment

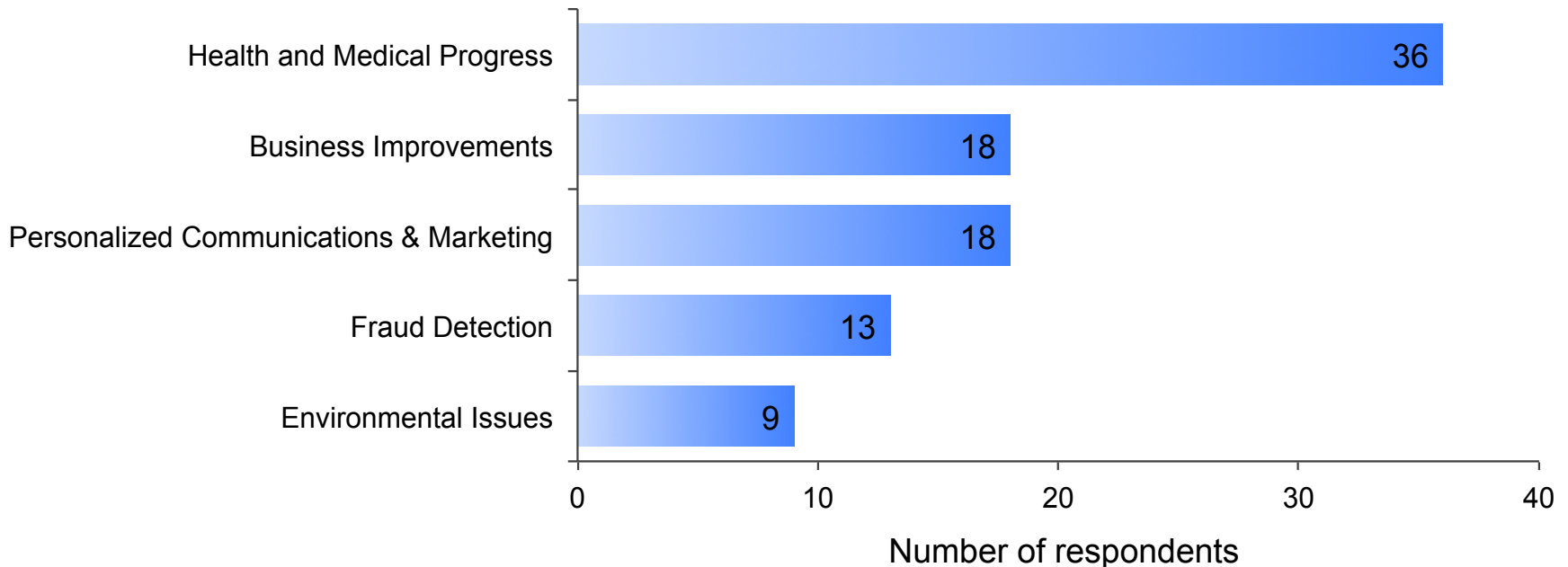
- Survey respondents shared their ideas about reasons for model non-deployment (an open-ended survey question).
- The largest number of respondents indicated that models are not deployed when the effort or cost to do so is too high.



Question: In cases where the results of your analytic projects are not deployed and/or utilized, what are the primary reasons that they are not used? (text box provided for response)

Positive Impact of Data Mining

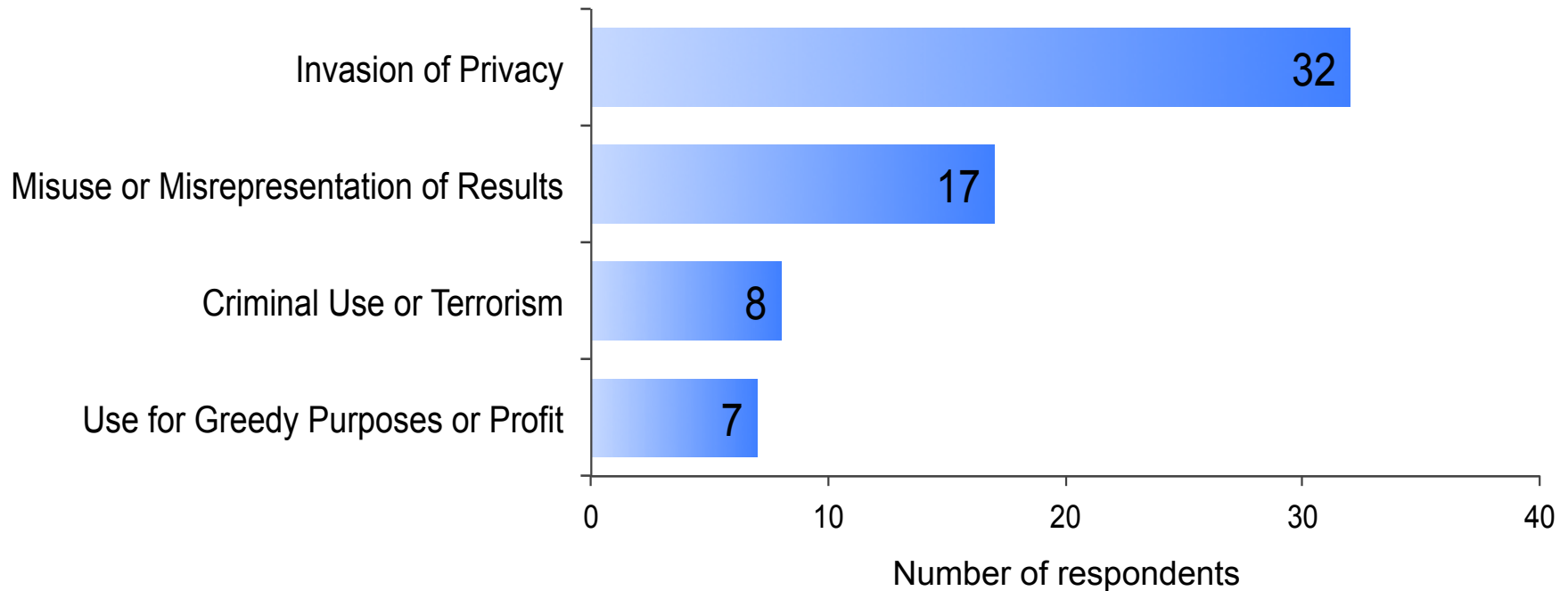
- Survey respondents shared their ideas about the positive impact of data mining on society (an open-ended survey question).
- The largest number of respondents identified positive impacts on our health and progress in medical fields.
- For a complete list of respondents' ideas about the positive impact of data mining, see www.RexerAnalytics.com/DMSurvey2011_PositiveImpact.



Question: Please share with us the best examples you know of that highlight the positive impact that data mining can have to benefit society, health, the world, etc. (text box provided for response)

Negative Impact of Data Mining

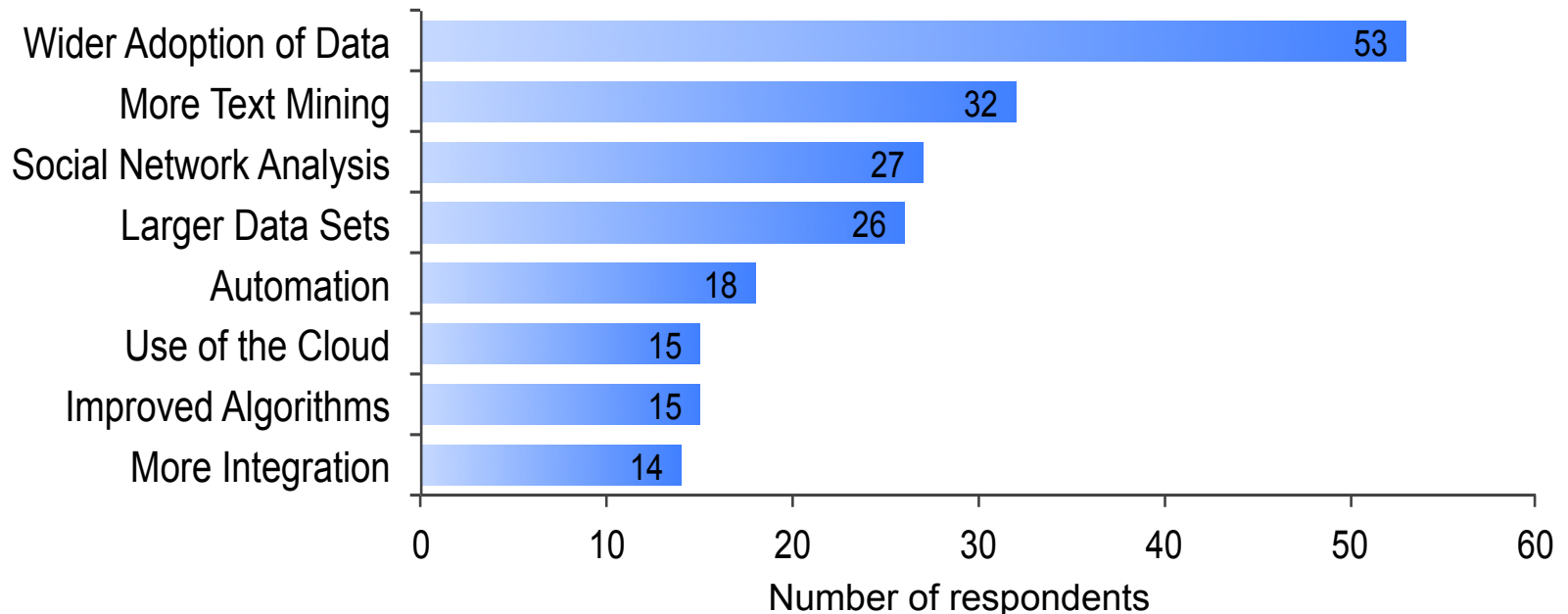
- Survey respondents also shared their ideas about the negative impact of data mining on society (an open-ended survey question).
- The largest number of respondents were concerned about the invasion of privacy that can sometimes accompany data mining.



Question: Please share with us the worst examples you know of that highlight a negative use of data mining. (text box provided for response)

Future Trends in Data Mining

- Survey respondents shared their ideas about future trends in data mining (an open-ended survey question).
- Many data miners think that there will be wider adoption of data mining in the future.
- Future visions of data mining are stable: the top three items are the same as last year.



Question: What do you envision as the primary future trends in data mining? (text box provided for response)

Appendix: Rexer Analytics

Rexer Analytics – Overview

Company Summary

- Small privately held consulting firm
- Founded in 2002
- Focus: Analytic and CRM Consulting (applied statistics & data mining)

Senior Staff

- Karl Rexer, PhD
- Paul Gearan
- Heather Allen, PhD
- Roberta Chicos

Example Projects

- Fraud detection
- Customer attrition analysis & prediction
- Text mining
- Customer segmentation
- Sales forecasting
- Market basket analysis
- Product allocation optimization
- CRM metric design & measurement
- Predictive models for campaign targeting & cross-sell
- Survey research (to understand customer needs & customer decision making)

Key Partners

- IBM (SPSS)
- Oracle
- Salford Systems
- Bennett Research
- Lincoln Peak
- Vlamis Software

Rexer Analytics – Clients

Additional clients were served. Some wish to remain anonymous, and others were served indirectly through partners.

2002

- CVS Pharmacy
- Fiserv
- Fleet Bank

2003

- CVS Pharmacy
- Fiserv
- Plymouth Bank
- Salford Systems

2004

- Hewlett Packard
- Quest Analytics (2 banking clients)
- Verizon
- New Direct

2005

- Hewlett Packard
- Quest Analytics (2 banking clients)
- ath Power (2 banking clients)
- Bridgewater State College
- DocSite

2006

- Hewlett Packard
- Quest Analytics (2 banking clients)
- ath Power (6 banking clients)
- Coverall
- Bridgewater State College
- Performance Programs
- Objective Management
- BBIQ
- Intellidyn

2007

- Hewlett Packard
- Quest Analytics (3 banking clients)
- ath Power (4 banking clients)
- Coverall
- Palladium
- Forbes Consulting
- Overture Networks
- Performance Programs
- Objective Management
- BBIQ

2008

- Hewlett Packard
- Quest Analytics (1 banking client)
- ath Power (4 banking clients)
- Coverall
- Palladium (9 clients)
- Forbes Consulting
- Raytheon
- Bernett Research (2 clients)
- Leader Networks
- One Day University
- Nexus Direct

2009

- Hewlett Packard
- Quest Analytics (1 banking client)
- ath Power (9 banking clients)
- Coverall
- Palladium (5 clients)
- Raytheon
- Bernett Research (4 clients)
- Leader Networks (3 clients)
- Accudata (2 clients)
- ITT Flow Control
- Stethographics
- MIT Epidemiology Group
- SNCR
- DLA Piper

2010

- Hewlett Packard
- Quest Analytics (1 banking client)
- ath Power (12 banking clients)
- Palladium (2 clients)
- Bernett Research (3 clients)
- Leader Networks
- Accudata (2 clients)
- ITT Flow Control
- SNCR
- DLA Piper
- Redbox
- ADT Security
- Lincoln Peak
- Loan Depot
- Oracle
- AboutFace
- Davol CR Bard

2011

- Quest Analytics (1 banking client)
- ath Power (8 banking clients)
- Objective Management
- Palladium (4 clients)
- Bernett Research (3 clients)
- Leader Networks (3 clients)
- Accudata (2 clients)
- SNCR
- Redbox
- ADT Security (2 divisions)
- Loan Depot
- Oracle
- MIT Epidemiology Group
- Meredith Corporation
- Shasta Partners
- HBO

Authors of the five Data Miner Surveys (2007-2011):

Heather Allen, PhD; Paul Gearan; & Karl Rexer, PhD

For more information contact:

Karl Rexer, PhD

krexer@RexerAnalytics.com

617-233-8185

Rexer Analytics

30 Vine Street

Winchester, MA 01890

USA

www.RexerAnalytics.com