CERTIFIED ANALYTICS PROFESSIONAL (CAP®)

EXAMINATION STUDY GUIDE

ACKNOWLEDGEMENTS

FOREWORD

As chair of the Analytics Certification Board, I congratulate the Study Guide committee on having assembled in short order such a comprehensive study guide for the Certified Analytics Professional (CAP®) program.

I know the guide is not going to satisfy everyone or directly provide them with answers for the test. It isn't designed to do so. It is designed to provide some information on central concepts embedded in the CAP program. It is up to the individual to determine his/her familiarity with the concept and decide whether more review or study on that topic is warranted.

The examination has 100 multiple choice test questions for each of which there is only one correct answer. The questions are both vendor and software neutral, designed to confirm that the test taker has the underlying knowledge necessary to know which steps to follow in an analytics process and to select the correct tools. The exam covers seven domains or areas of analytics practice: business problem framing, analytics problem framing, data, methodology (approach) selection, model building, deployment, and model life cycle management. A sample of the type of questions is available with this guide and can also be accessed through the Candidate Handbook. These sample questions will never appear on an exam. Each sample gives not only the correct answer but also provides rationale for why each is (in)correct.

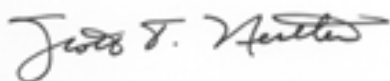What are individual benefits of the Certified Analytics Professional program?

- Advances your career by setting you apart from the competition
- Promotes personal satisfaction from accomplishing a key career goal
- Helps improve your overall job performance by setting you on a course for continual professional development
- Recognizes the investment you have made in your career
- Demonstrates commitment to the field

What are employer benefits of the Certified Analytics Professional program?

- Helps with identifying and developing qualified analytics professionals
- Proves to stakeholders that your organization follows industry-standard analytics practice
- Provides a career path to encourage employees
- Useful as positive factor in responding to proposals
- Indicates a company willing to invest in its employees
- Indicates a willingness to maintain up to date knowledge

You are to be applauded for seeking certification. While the exam is the most pressing hurdle to achieving the CAP, it is not the only criterion. The Certified Analytics Professional program depends on each of the Five E's: They are adherence to the Code of Ethics, Effective mastering of soft skills, acceptable levels of Experience and Education and finally, successfully passing the Exam. The result of this program is a well-rounded analytics professional who can work in many fields to provide analytic leadership and support.

The Analytics Certification Board wishes all candidates complete success in their certification process. If I, or they, can be of help, feel free to contact me at acb@informs.org or email our Certification Manager at info@certifiedanalytics.org.

Scott Nestler, US Army
Chair Analytics Certification Board

TABLE OF CONTENTS

TABLE OF CONTENTS

List of Figures

Figure 1: Possible shapes of analytics knowledge (*OR/MS Today*, June 2013)

Figure 2: Kano's Requirements Model (used under Creative Commons, http://creativecommons.org/licenses/by-nc-nd/3.0/us/)

Figure 3: Input Table by Alan Taber, CAP® (used with permission)

Figure 4: Black Box Sketch by Alan Taber, CAP® (used with permission)

Figure 5: INFORMS CAP® Methodology Classification (used with permission)

Figure 6: Sample software application characteristics by Rami Musa, CAP® (used with permission)

* Tasks performed by analytics professionals beyond CAP® certification level

The Institute for Operations Research and the Management Sciences (INFORMS) is an international scientific society with more than 11,000 members, including Nobel Prize laureates, dedicated to applying scientific methods to help improve decision making, management, and operations. Members of INFORMS work in business, government, and academia.

INFORMS serves the scientific and professional needs of operations research analysts, experts in analytics, consultants, scientists, students, educators, and managers, as well as their institutions, by publishing a variety of journals that describe the latest research in operations research.

INFORMS commissioned a study in 2010 by Capgemini Consulting to evaluate the need for embracing the analytics community as a key membership strategy for the Institute. The Capgemini study concluded that INFORMS should embrace the analytics community and that one of the first initiatives should be a detailed study on the development of a certification and training program to meet the needs of this market.

Further market research corroborated this finding. Mike Hamm of Michael Hamm and Associates wrote of his market research—done in 2011 on behalf of INFORMS—that "I have never seen a candidate audience where [there is such] a high degree of political interest regarding the potential composition and architecture of a future certification program. Everybody wants a piece of this... ."

INFORMS defines analytics as the scientific process of transforming data into insight for making better decisions. It is seen as an end-to-end process beginning with identifying the business problem to evaluating and drawing conclusions about the prescribed solution arrived at through the use of analytics. Analytics professionals are skilled at this process.

INFORMS established the analytics certification program to advance the use of analytics by setting agreed upon standards for the profession. The program also advances the analytics profession by providing a means for organizations to identify and develop qualified analytics professionals, by contributing to the career success and continued competence for analytics professionals, and by improving the credibility and visibility of the analytics profession.

INFORMS vision for the Certified Analytics Professional (CAP®) program is to advance the use of analytics to transform the world by setting agreed-upon standards for the profession. The INFORMS mission for the CAP® program is

to advance the analytics profession by providing a high-quality program of certification and by promoting continuing competence for practitioners.

Once INFORMS decided to pursue a certification program, the practicalities of creating the program and its accompanying exam were addressed. According to an article in *Analytics Magazine*, September/October 2012 by Scott Nestler, Jack Levis, and Bill Klimack, the CAP® program is appropriate for the analytical semi-professionals as well as the analytical professionals. However, it will not be a suitable certification for the "analytical amateurs" as depicted in the following graphic (Nestler et al. 2012, Figure 2). The assessment instrument—the exam—contains 100 multiple choice test items and is being administered via paper and pencil for the first year. INFORMS is investigating the possibility of moving to computer-based testing for subsequent years to facilitate serving its international membership.

	(a)	(b)	(c)	(d)	(e)
Analyst Type (Davenport)	Amateur	Semi-professional	Professional	Professional	Professional
Shape	"bar-shaped"	"bar-shaped"	"I-shaped"	"T-shaped"	"Pi-shaped"
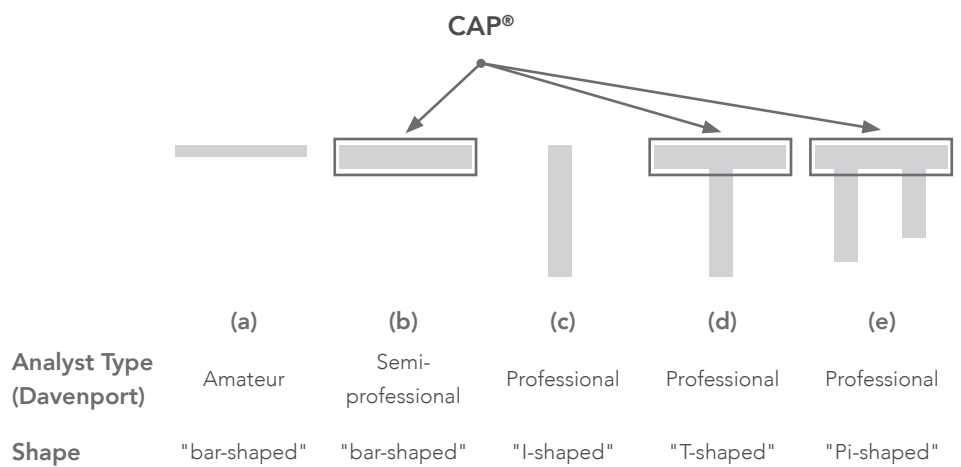
Figure 1: Possible shapes of analytics knowledge

For more information on the development of the CAP program, read "Steering Toward Analytics" in *OR/MS Today*, June 2013 (p. 30) by Gary Bennett and Jack Levis.

A job task analysis (JTA) is a comprehensive description of the duties and responsibilities of a profession, occupation, or specialty area; our approach consists of four elements: 1) domains of practice, 2) tasks performed, 3) knowledge required for effective performance on the job, and 4) domain weights that account for the importance of and frequency with which the tasks are performed. More specifically, the JTA for the CAP® program can be viewed as an outline of a partial body of knowledge, as it represents a delineation of common or typical tasks performed and knowledge applied by analytics professionals, grouped together in a hierarchical domain structure. In the course of analytics work, these tasks may be performed multiple times with modifications based on data, findings, and results, as part of ongoing feedback loops that are routinely a part of practice. The JTA serves as the test blueprint for exam development and links what is done on the job to what is measured by the certification examination. This linkage is necessary to establish a valid, practice-related examination. It is important to realize that the JTA is a dynamic document that will change in the future to reflect best practices and changes in the analytics profession.

The JTA study defines the current knowledge, skills, and abilities (KSAs) that must be demonstrated by analytics professionals to effectively and successfully provide these services. KSAs are validated according to their frequency of use and importance. The JTA also serves as a "blueprint" for the content (performance domains) of the INFORMS CAP® examination.

INFORMS upholds stringent guidelines for the construction and implementation of the examination development and administration process. An 11-member panel of subject matter experts (SMEs) was selected to develop the first JTA for the CAP® credential. This group was called the Analytics Certification Job Task Analysis Working Group.

The following leaders in the analytics profession were selected to participate in this important project:

- Arnold Greenland (IBM Global Business Services)
- Bill Klimack (Chevron)
- Jack Levis (UPS)
- Daymond Ling (Canadian Imperial Bank of Commerce)
- Freeman Marvin (Innovative Decisions, Inc.)
- Scott Nestler (Naval Postgraduate School)
- Jerry Oglesby (SAS)
- Michael Rappa (North Carolina State/ Institute for Advanced Analytics)
- Tim Rey (Dow Chemical)
- Rita Sallam (Gartner)
- Sam Savage (Stanford/Vector Economics)

The findings of this working group were then validated by a random sample of practicing analytics professionals. Feedback from this survey resulted in slight modifications of the performance domains, tasks, and knowledge that comprise the test blueprint that determines the content of the CAP® examination.

In developing the JTA, members of the working group relied on their knowledge of practice gained from years of experience, academic program content, corporate job descriptions in analytics, and articles from professional and scholarly publications.

The following table includes the final domains and weights derived from the JTA and a review of validation survey recommendations.

Domain	Approximate Weight
I. Business Problem (Question) Framing	12%–18%
II. Analytics Problem Framing	14%–20%
III. Data	18%–26%
IV. Methodology (Approach) Selection	12%–18%
V. Model Building	13%–19%
VI. Deployment	7%–11%
VII. Model Life Cycle Management	4%–8%

The INFORMS CAP® examination is based on the following test blueprint derived from the JTA process. The final agreed-upon weights reflect the percentage of questions from each domain that will be included in each test form.

The JTA and the test blueprint resulting from this process will be reviewed periodically and updated as needed to reflect current practices in analytics. The list of domains and key tasks follows:

(12%–18%) Domain I Business Problem (Question) Framing

(The ability to understand a business problem and determine whether the problem is amenable to an analytics solution.)

    T-1 Obtain or receive problem statement and usability requirements
    T-2 Identify stakeholders
    T-3 Determine whether the problem is amenable to an analytics solution
    T-4 Refine the problem statement and delineate constraints

T-5 Define an initial set of business benefits

T-6 Obtain stakeholder agreement on the business problem statement

(14%–20%) Domain II Analytics Problem Framing

(The ability to reformulate a business problem into an analytics problem with a potential analytics solution.)

T-1 Reformulate problem statement as an analytics problem

T-2 Develop a proposed set of drivers and relationships to outputs

T-3 State the set of assumptions related to the problem

T-4 Define key metrics of success

T-5 Obtain stakeholder agreement on the approach

(18%–26%) Domain III Data

(The ability to work effectively with data to help identify potential relationships that will lead to refinement of the business and analytics problem.)

T-1 Identify and prioritize data needs and sources

T-2 Acquire data

T-3 Harmonize, rescale, clean, and share data

T-4 Identify relationships in the data

T-5 Document and report findings (e.g., insights, results, business performance)

T-6 Refine the business and analytics problem statements

(12%–18%) Domain IV Methodology (Approach) Selection

(The ability to identify and select potential approaches for solving the business problem.)

T-1 Identify available problem solving approaches (methods)

T-2 Select software tools

T-3 Test approaches (methods)*

T-4 Select approaches (methods)*

(13%–19%) Domain V Model Building

(The ability to identify and build effective model structures to help solve the business problem.)

T-1 Identify model structures*

T-2 Run and evaluate the models

T-3 Calibrate models and data*

T-4 Integrate the models*

T-5 Document and communicate findings (including assumptions, limitations, and constraints)

*Tasks that are beyond the scope of the CAP® certification exam and will not be tested.

(7%–11%) Domain VI Deployment

(The ability to deploy the selected model to help solve the business problem.)

    T-1 Perform business validation of the model
    T-2 Deliver report with findings; or
    T-3 Create model, usability, and system requirements for production
    T-4 Deliver production model/system*
    T-5 Support deployment

(4%–8%) Domain VII Model Life Cycle Management

(The ability to manage the model life cycle to evaluate business benefit of the model over time.)

    T-1 Document initial structure
    T-2 Track model quality
    T-3 Recalibrate and maintain the model*
    T-4 Support training activities
    T-5 Evaluate the business benefit of the model over time

The knowledge statements for the CAP® program have been identified but not individually assigned to each task. The knowledge statements appropriate to a given task have been used. Not all statements are appropriate for all tasks, although there may appear to be some blanks in coverage this is not the case.

K-1   Characteristics of a business problem statement (i.e., a clear and concise statement of the problem describing the situation and stating the desired end state or goal)

K-2   Interviewing (questioning) techniques (i.e., the process by which a practitioner elicits information and understanding from business experts, including strategies for the success of the project)

K-3   Client business processes (i.e., the processes used by the client or project sponsor that are related to the problem)

K-4   Client and client-related organizational structures

K-5   Modeling options (i.e., the analytic approaches available for seeking a solution to the problem or answer to the question including optimization, simulation, forecasting, statistical analysis, data mining, machine learning, etc.)

K-6   Resources necessary for analytics solutions (e.g., human, data, computing, software)

K-7   Performance measurement (i.e., the technical and business metrics by which the client and the analyst measure the success of the project)

K-8   Risk/return (i.e., trade-offs between prioritizing the primary objective and minimizing the likelihood of significant penalty taking into account therisk attitude of the decision maker)

K-9   Presentation techniques (i.e., strategies for communicating analytics problems and solutions to a broad audience of business clients)

K-10   Structure of decisions (e.g., influence diagrams, decision trees, system structures)

K-11   Negotiation techniques (i.e., strategies and methods that allow the analytics professional to reach a shared understanding with the client)

K-12   Data rules (e.g., privacy, intellectual property, security, governance, copyright, sharing)

K-13   Data architecture (i.e., a description of how data are processed, stored, and used in organizational systems including conceptual, logical, and physical aspects)

K-14   Data extraction technologies (e.g., scripting, spreadsheets/databases,connection tools, standards-based connectivity options, unstructured data extraction tools)

K-15   Visualization techniques (i.e., any technique for creating images, diagrams or animations to communicate a message including data visualization, information visualization, statistical graphics, presentation graphics, etc.)

K-16   Statistics (descriptive, correlation, regression, etc.)

K-17   Software tools

The five E's are ethics, education, experience, examination, and effectiveness. These are the five pillars of the Certified Analytics Professional.

The CAP® credentialed person will have read, agreed to, and signed the code of ethics that governs behavior of a professional analyst. This code was created by the Task Force who are among the originators of the program (see Figure 1). The code is intended to describe the accepted behavior of an analytics professional. All candidates for the CAP® must agree to the code of ethics as part of the application process. Actions that are opposed to the code of ethics may be reason to rescind the CAP® credential.

Education is considered essential for the analytics professional. Candidates must have at least a bachelor's degree from a regionally accredited college or university. Experience goes hand in hand with education as part of the prerequisites for application. The higher and more appropriate the education earned, the less experience is required.

Examination is the fourth leg or pillar of the CAP® program. Through examination we seek confirmation that the applicant has knowledge of those areas of the job/task analysis that are considered essential for practice. Because the examination is based on a broad spectrum of practice rather than the content of a course or series of courses, it must be constructed with due care. Each test item or question has been created carefully so as to ensure a fair, valid, and reliable examination that discriminates against no one except for those who do not have the knowledge to earn the CAP® credential. Each item is reviewed and refined numerous times by a committee of subject matter experts in the field of analytics. The sole reason to use a test item is as a tool to determine who is knowledgeable. Because there may be a lot riding on the successful completion of the exam, the test items must be carefully crafted.

All test items are written with reference to the specific domain, task, and knowledge statements outlined earlier. Test items are also sourced to ensure that all items are readily available and should be known to everyone who is an analytics professional. No items are written based on proprietary data or sources that are known only to a select few. For examples, see the Candidate Handbook that contains 24 questions or items that are indicative of the style of test item but that do not themselves appear on the exam. In the future, there may be additional items that we will release from the item bank and use as practice test questions. The CAP® program is so new that INFORMS does not yet have items that have outlasted their usefulness as a discriminatory tool to distinguish between the knowledgeable and those who do not yet possess the knowledge.

The rules for item writing are specific and few:

- Avoid negative stems or questions as much as possible

- Do not use 'All of the above' or 'None of the above' as answer options

- Avoid excess verbiage

- Avoid disadvantaging any part of the test population but the unknowing

- Ask only one question at a time

- Ensure that the incorrect answers are incorrect for a specific reason

Effectiveness is the art of applying your knowledge and skill in a way that enables achievement of your organization's goals. The soft skills required are dealt with more fully in Appendix A: Soft Skills. Nevertheless, the skilled analyst must be diplomatic and aware enough to understand the context of the business problem and the stakeholder agendas involved while not allowing that understanding to bias the process or the truth thereby developed.

The Certified Analytics Professional (CAP®) program is not the work of one person or one department: it would not have been possible without the support of professionals in the field. You can see a long list of those professionals on the INFORMS website under Contributors (www.informs.org/Certification-Continuing-Ed/Analytics-Certification/Contributors).

This study guide is the culmination of a massive collaborative effort among concerned professionals to develop a guide that will assist future CAPs. The guide is not intended to give the answer to each and every test question. Rather, it is intended to guide the individual toward the knowledge of an analytics professional and to let the individual use his or her discretion as to areas that warrant further study. The guide includes reference materials for this further study. The guide is also intended to be a comprehensive outline for those who are working in, intend to work in, or are preparing to work in an analytics area. Because being an effective analytics professional is as much of an art as a science, the study guide relies heavily on case studies, examples, and stories.

If you have comments on the guide, the certification program, or wish to assist with the further development and dissemination of the CAP® program, please feel free to e-mail certification@informs.org.

WHAT WILL YOU LEARN IN THIS CHAPTER?

In this chapter, you will learn about the first step of an analytics project: framing the business problem. You will learn, as a part of these processes, how to determine the business problem, identify and enlist stakeholders, determine if the problem has an analytics solution, refine the problem statement as necessary, and define the set of business benefits.

Learning Objectives

1. Obtain or receive the problem statement and usability requirements

2. Identify stakeholders

3. Determine whether the problem is amenable to an analytics solution

4. Refine the problem statement and delineate constraints

5. Define an initial set of business benefits

6. Obtain stakeholder agreement on the business problem statement

Key Concepts/Fundamentals

OBJECTIVE 1. RECEIVE & REFINE THE BUSINESS PROBLEM

A business problem statement generally starts by describing a business opportunity or threat, or an issue in broad terms. For example, it could simply start by saying 'our growth has been stagnant for the last two years' or a bit less broad 'our Seattle plant is experiencing production problems and is missing deadlines.' Most client firms in their early meetings with you (the analytics professional) will tend to report what they are experiencing as problems. As they do that, they will use their own language and key terms. Do get definitions of all terms, as meanings change between organizations.

Another factor to consider is that the client firm representatives in these meetings also play an important role in what is reported and how it is reported. It is natural that each representative (of the firm) uses their own lenses and contexts to report (and thus frame) the way they see the problem. These views are all very important on their own merits because they inform the analyst in some useful way. However,

because of the individual lenses used to report these observations, sometimes these views can have a good degree of variance regarding causes and effects, and thus may obscure the real issues.

One popular way to frame a business opportunity or problem is to obtain reliable information on the five W's: who, what, where, when, and why.

- Who: are the stakeholders who satisfy one or more of the following with respect to the project: funding, using, creating, or affected by the project's outcome.

- What: problem/function is the project meant to solve/perform?

- Where: does the problem occur? Or where does the function need to be performed? Are the physical and spatial characteristics articulated?

- When: does the problem occur, or function need to be performed? When does the project need to be completed?

- Why: does the problem occur, or function need to occur?

OBJECTIVE 2. IDENTIFY STAKEHOLDERS

Of the five W's, who (the stakeholders are) is probably the most critical to the long term success of the project. Stakeholders are anyone affected by the project, not just those in the initial meetings, and they may have different levels of input or involvement during the project. A stakeholder analysis helps identify the following:

- The interests of all stakeholders, who may affect or be affected by the project, along with their constraints.

- Potential issues that could disrupt the project.

- Key people for information distribution during execution phase.

- Groups that should be encouraged to participate in different stages of the project.

- Communication planning and stakeholder management strategies during the project planning phase.

- Ways to reduce potential negative impacts and manage negative stakeholders.

OBJECTIVE 3. DETERMINE WHETHER THE PROBLEM IS AMENABLE TO AN ANALYTICS SOLUTION

Before more time and money is spent on solving the problem, it is time to figure out if this problem is likely to have an analytics solution. First of all, does the answer and the change process to get there lie within the organization's control? Second, does the requisite data exist or can it be obtained? Third, can the likely problem be solved and/or modeled? Last, but perhaps most importantly, can the organization accept and deploy the answer? The problem may not be amenable to an analytics solution because of the characteristics of the problem or the limitations of the analytic tools/methods available. The problem statement could be reassessed to make it amenable to the available analytic tools/methods, or if this is not possible, the project deemed not feasible. If there isn't a feasible way forward, the ethical analyst will say so to the key stakeholders.

For the Seattle plant example, it may be decided to use mathematical optimization software to improve the plant's process. This will work as long as data exist on inputs and outputs for each step in the plant process, and as long as the stakeholders are willing to accept new ways of operating that won't necessarily match current work policies and procedures.

OBJECTIVE 4. REFINE PROBLEM STATEMENT & DELINEATE CONSTRAINTS

After the initial analysis, it may be necessary to refine the problem statement to make it more accurate, more appropriate to the stakeholders, or more amenable to available analytic tools/methods. As part of this process, it will become necessary to define what constraints the project will operate under. These constraints could be analytical, financial, or political in nature.

For the Seattle plant example, an optimization problem with a large number of constraints or a complex objective function may not be solvable within the capability of the available software/hardware combination. In this case the problem may need to be restated with fewer constraints and/or a less complex objective function. This may cause the problem statement to be updated to make sure that the approach will satisfy—just to name a few of the potential constraints—desired accuracy and repeatability, program cost, timeframe, and number of stakeholders impacted, either positively or negatively.

OBJECTIVE 5. DEFINE AN INITIAL SET OF BUSINESS BENEFITS

With the problem statement set, it is now possible to define the initial set of business benefits. These benefits may be determined quantitatively or qualitatively. If quantitative, it may be financial (e.g., net present value) or contractual (e.g., service level agreements). This is also known as the business case.

For the Seattle plant example, an initial determination of the financial benefit due to optimal use of resources should be determined along with an initial view of the required project goals determined, e.g., plant is currently losing money at the rate of 3% of gross sales with current performance and needs to come to 5% margin on gross sales. The key profit driver is on-time performance, which is currently 68% and needs to get to 98%. How will it get there? At this stage we think it is because there is plant capacity being wasted, so we're going to look at optimizing our scheduling and manufacturing processes to reduce overall time by reducing queue and wait time. You'll note that we haven't said, yet, that we're going to simulate incoming orders with one distribution and performance of each machine on the floor with their own distributions, even though we may be thinking about doing just that. At this stage, the problem is a business problem and the objectives are business objectives.

OBJECTIVE 6. OBTAIN STAKEHOLDER AGREEMENT ON THE PROBLEM STATEMENT

With the problem statement refined and the initial business benefits determined, it is necessary to obtain stakeholder agreement before proceeding further with the project. It may be necessary to repeat this cycle several times until stakeholder concurrence with the particulars of the project are achieved and permission to proceed is granted. At the end of this process, you will have agreement on the project's objectives, initial approach, and resources to get there.

SUMMARY

Although business problem framing is not the analytical heart of an analytics project, it is probably the most important because it sets the expectations and limitations of the project.

Davenport T, Kim J (2013) *Keeping up with the Quants: Your Guide to Understanding and Using Analytics* (Harvard Business Review Press, Boston).

Framing the problem at https://www.boundless.com/business/management/ decision-making/observation-framing-the-problem/.

Kirkwood CW (1997) *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets* (Duxbury Press, Pacific Grove, CA).

Lindstrom C (2009) How to write a problem statement, March 18, http://www. ceptara.com/blog/how-to-write-problem-statement.

Nixon NW (2013) Focus first on framing, not solving, the problem, April 18, http:// philadelphia.regionsbusiness.com/print-edition-commentary/focus-first-on- framing-not-solving-the-problem/.

Seelig T (2013) Shift your lens: The power of re-framing problems. Seelig T, ed. *inGenius: A Crash Course on Creativity* (HarperOne, New York), http://stvp. stanford.edu/blog/?p=6435.

Spradlin D (2012) The power of defining the problem, September 25, http:// blogs.hbr.org/cs/2012/09/the_power_of_defining_the_prob.html.

THIS PAGE IS DELIBERATELY LEFT BLANK

WHAT WILL YOU LEARN IN THIS CHAPTER?

This chapter is all about the dialogue between the business people who have a problem that they need to solve and the analytics folks who will give them the information required to solve the problem. This dialogue is mediated by the analytics professional (YOU) who is trusted by both sides because you are fluent in the language and culture of each side. As with any translation effort between two different groups, much of what follows are simple precepts to keep the sense of the business problem while decomposing it into actionable analytics pieces.

Learning Objectives

1. Reformulate a problem statement as an analytics problem

2. Develop a proposed set of drivers and relationships to inputs

3. State the set of assumptions related to the problem

4. Define key metrics of success

5. Obtain stakeholder agreement on the approach

Key Concepts/Fundamentals

OBJECTIVE 1. REFORMULATING THE BUSINESS PROBLEM STATEMENT AS AN ANALYTICS PROBLEM

There's an apocryphal story of a Black & Decker sales convention. The VP of sales gets up to the dais, and says, "Folks, I have some bad news for you. We've done some detailed customer surveys to find out what our customers care about. They couldn't care less about our carbide tips, or the voltage rating of our drills. In fact, they'd rather not think about drills at all! What our customers want is to hang a picture, or put up drywall, or do any number of other jobs. Our job is to help them do just that." Similarly, your business and operational stakeholders likely could not care less about **how** you and your team are going to solve their problem. They just want it to be solved reliably and deliver the results.

The first step is to decode the business problem statement to get to the analytics problem. There are many ways to do this, some more formal than others. In simple

terms, you are translating the "what" of the business problem into the "how" of the analytics problem.

1. What result do we want?

2. Who will act?

3. What will they do?

4. What will change in the organization as a result of the new information generated?

For example, a company wishes to increase market share, but what is the underlying problem they need to address? Are they, for instance, emphasizing carbide-tipped drills to someone who only wants to hang a picture?

One formal method of decomposition is quality function deployment (QFD) (http://www.ieee.li/tmc/quality_function_deployment.pdf). This is a rigorous process that maps the translation of requirements from one level to the next, e.g., from the business level to the first analytics level, from the first analytics level to the second level, etc.

Whether you are formally decomposing and parsing a complex business statement, or you are less formally brainstorming with a project sponsor, it is critically important to account for tacit as well as formal requirements. The best known model in this area is Kano's requirements model (Figure 2). It distinguishes between unexpected customer delights, known customer requirements, and customer must-haves that are not explicitly stated.

CUSTOMER SATISFACTION

exciting requirements

normal requirements

DON'T FULFILL EXPECTATIONS

DO FULFILL EXPECTATIONS

expected requirements

CUSTOMER DISSATISFACTION

Figure 2: Kano's Requirements Model (used under Creative Commons, http://creativecommons.org/licenses/by-nc-nd/3.0/us/)
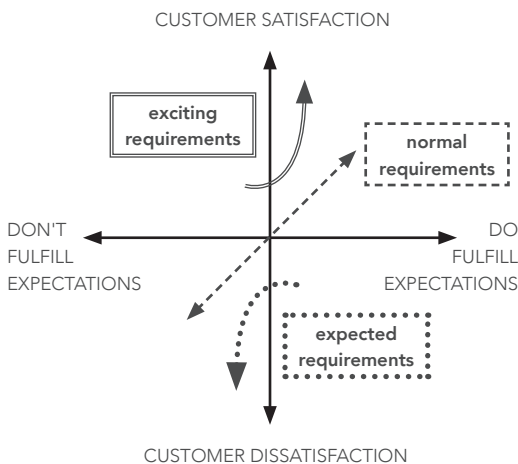
Often there are business or operational requirements that are taken for granted by those stakeholders that if not surfaced will result in customer dissatisfaction, particularly items that come under the heading of "that's the way we always do things." Now, there are times when those requirements or assumptions need to be challenged, but they can't be challenged until they are brought to light. When you ask your business stakeholders for a list of what requirements they have, they will tend to focus on the "normal

requirements," not the "expected requirements." As the analytics professional charged with translating business requirements into the problem statement, you really need to probe to make sure that you have the entire appropriate context as well, including the expected requirements.

OBJECTIVE 2. DEVELOP A PROPOSED SET OF DRIVERS & RELATIONSHIPS TO OUTPUTS

These next three items are related. Your input/output functions are strongly related to your assumptions about what is important about this problem as well as the key metrics by which you'll measure the organizational response to the problem.

We'll start by defining the input/output functions of the problem at hand. As with any of these areas, you can be as formal or informal as you like, but sketches and diagrams certainly help communicate with your stakeholders and help get everyone on the same page.

Here's a very simple example: An organization wants to predict the number of detected software defects over the next six months. That's the output. The inputs would be elicited from stakeholder interviews, using questions like, "What future activities will add to our rate from where we are today?", "What will decrease our rate from where we are today?", "Will we add interfaces or components to the testing?", "Will we materially change the size of the test team?", etc. Bear in mind that you aren't looking for causation at this stage, just ideas around which you'll form some hypotheses against which you'll test your model later.

Once you have these inputs and a general sense of their predicted effects, you have a choice of how to communicate them to the team at large. A simple table (Figure 3) is one approach. A black box sketch (Figure 4) is another approach. How you do it isn't nearly as important as doing it in a way that the people you're working with will understand.

INCREASING FACTORS		DECREASING FACTORS	
NAME	SCALE	NAME	SCALE
NEW INTERFACES	LESS THAN 1	TEST TEAM SIZE	LESS THAN 1
CUSTOMER SITE DEPLOYMENT	1-10	TIME SINCE LAST NEW FUNCTIONALITY	LESS THAN 1
...	...	...	...

Figure 3: Input Table

Figure 4: Black Box Sketch

Even these simple examples help illustrate the concept. The idea here is to make the inputs visible and start getting agreement among the team on the direction and scale of the relationships to bound the problem and to create the related hypotheses that you'll use later to attack the data. A point you'll want to emphasize to the team is that these are preliminary assumptions and while your best estimate is needed, it is still just an estimate and is subject to change depending on what reality turns out to be. The danger we're trying to avoid here is what Kahneman calls "anchoring." People have a tendency to hang on to views that they've seen and held before, even if they are incorrect. Reminding them that these are initial and preliminary, rather than finalized views, helps mitigate the anchoring effect.

OBJECTIVE 3. STATE THE SET OF ASSUMPTIONS RELATED TO THE PROBLEM

This is where you set the boundaries of the problem. As you look at your input drivers, each likely has one or more assumptions embedded in it that needs to be surfaced and listed. Additionally, some complexities can be trimmed away if their presumed effect on the answer is less than the effort required to handle them.

As Stephen R. Covey (2004, p. 24) said, "We simply assume that the way we see things is the way they really are or the way they should be. And our attitudes and behaviors grow out of these assumptions." Common practice assumptions in your organization also need to be listed and questioned regularly to ensure that they are either still valid or that the problem statement needs to change to incorporate changes to them.

OBJECTIVE 4. DEFINE THE KEY METRICS OF SUCCESS

There is a truism quoted by many people that "what is measured, improves" (cf. Drucker, Pearson's Law, Hawthorne effect). This ties directly to the business problem statement, but goes down one level further to the items that comprise the key success metric. For example, if the business problem is that the organization wants to increase return on sales from 10% to 12%, you might decompose that a few different ways. One way is to give each business group that goal, or even to give each group the objective of reaching 13%, figuring that some won't make it and on average we'll be okay. Another way is to look at your value chain and give each group a target: cost of goods reduction of five percentage points, general and administrative reduction of three percentage points, etc. These metrics need to be negotiated, published, committed to, and tracked so that your team knows where you are and what to do next. As is the case throughout this chapter, you have to make sure that all facets of the business problem are incorporated in the metrics. After all, if you don't say how something will be measured, you don't know how you're doing, and you can't succeed.

OBJECTIVE 5. OBTAIN STAKEHOLDER AGREEMENT

Although you've been in touch with your business stakeholders at some level all along, this is when you come back to them to walk them through your assumptions and approach and what the final answer will look like to be sure that you really are answering the business problem. Whether in the form of a formal presentation, you want your assumptions acknowledged along with the reframing you did from the business problem, and the key metrics you will be using to mark progress toward the solution.

Many people tend to think of stakeholders as people in positions "above" the analytics team. It is true that there is a group of stakeholders that are the ones with the business need and who are paying for the effort. But just as importantly, you must also have an agreement with the people executing the analytics work that your methods and hypotheses are workable in the time and budget allotted to get the work done.

The output of this stakeholder agreement will vary by organization, but should include the budget, timeline, interim milestones (if any), goals, and any known effort that is excluded as out of scope. The key is to get all the pieces we've noted in this chapter verbally discussed, documented, and visibly agreed to by all parties. It can be tempting to settle for e-mails or written documents only and desk-side reviews. For all but the simplest problems, this is a mistake. Translation of problems from the business domain to the analytics domain, or truly from any

given domain to another domain, requires that all parties agree to definitions and terms, which really does require full and frank discussion. Otherwise, errors will creep in and what was delivered will miss critical unstated requirements. If you allow your project to rely on written communication only, you've missed the opportunity to correct misapprehensions when it is still cheap to do so.

SUMMARY OF KEY TERMS

Decomposition: the act of breaking down a higher-level requirement to multiple lower-level requirements (http://www.hq.nasa.gov/office/codeq/software/ComplexElectronics/l_requirements2.htm).

Requirements: a requirement should be unitary (no conjunctions such as and, but, or or), positive, and testable.

SUMMARY

Faithful translation of the business problem statement into an analytics problem statement requires the following:

- Understanding the business case for solving this particular problem

Framing the business case as an actionable analytics problem by:

- Defining the key input and output drivers

- Surfacing and understanding individual and organizational assumptions

- Assigning goals to each sub-group affected by the problem

Full and frank review of the approach with the business stakeholders and the analysts to ensure that the problem can be attacked as planned and that a successful attack will yield the desired business result.

FURTHER READING

Albright SC, Winston W, Zappe C (2011) *Data Analysis and Decision Making*, 4th ed. (South-Western Cengage Learning, Mason, OH).

Covey S (2004) *The 7 Habits of Highly Effective People* (Simon & Schuster, New York).

Crow KA (1992) Quality Function Deployment, http://www.ieee.li/tmc/quality_function_deployment.pdf.

National Aeronautics and Space Administration (2009), Assurance Process for Complex Electronics, http://www.hq.nasa.gov/office/codeq/software/ComplexElectronics/l_requirements2.htm.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

THIS PAGE IS DELIBERATELY LEFT BLANK

WHAT WILL YOU LEARN IN THIS CHAPTER?

Analytics is defined by INFORMS as the scientific process of transforming data into insight for making better decisions.   In this section we will see how data collection, manipulation and analysis support the analytic framework from problem identification to model building and management. Data transformation starts with data element definition and potential source identification. Once sources are identified, collection of new data and extraction and transformation of existing data can begin. Often data will need to be cleaned to address incorrect and/or missing data points. Finally, the data must be properly formatted for use in the common database and loaded into it.

Learning Objectives

By the end of this chapter, you should be able to

1. Identify and prioritize data needs and resources

2. Identify means of data collection and acquisition

3. Determine how and why to harmonize, rescale, clean and share data

4. Identify ways of discovering relationships in the data

5. Determine the documentation and reporting of findings

6. Use data analysis results to refine business and analytics problem statements

Key Concepts/Fundamentals

OBJECTIVE 1. IDENTIFY & PRIORITIZE DATA NEEDS & RESOURCES

Data reduces our uncertainty about the values assigned to variables of interest in the analysis.

Analysis typically uses `hard data', i.e., data that is obtained by scientific observation and measurement (e.g., experimentation). But much of our information is frequently soft, e.g., gleaned from interviews and reflective opinions and preferences. Hence it will be important to convert this soft information into scientific data. The

traditional way in which soft data is converted into hard data is to hypothesize an artificial individual whose preferences and beliefs can be completely described with hard data. (In economics, this artificial individual is called the `economic man' and is viewed as totally rational.) We then determine what hard data would be required so that this artificial individual's behavior coincides with that of the actual individual with soft data. We then solve the analytical problem as if our actual individual could be described by this artificial individual.

Probably the most successful example of this approach is conjoint measurement or analysis which posits that the behavior of the actual individual can be described by an artificial individual whose preferences are described by a utility function. The utility function for various outcomes is first specified as a parametric function of observable attributes of that item. If this utility function were known, then it is, in theory, straightforward to specify which of several items an individual would choose or how an individual would rank several items. To determine the parameters of this utility function, individuals are then asked to either specify which, of several hypothetical alternatives, they prefer or how they would rank different items in order of preferability. The parameters are then calibrated so as to minimize the disparity between what the individual actual prefers and what the model predicts the individual should prefer.

Other methods analogous to conjoint have been developed when uncertainty is involved. For example, it may be necessary to develop a `subjective probability' as a summary measure of an individual's beliefs about whether an event occurs. To assess an individual's subjective probability, consider a random mechanism (e.g., a roulette wheel in a casino or a table of random numbers). For any frequency f between zero and one, this random mechanism can be used to define an uncertain event $A(f)$ which occurs with probability f. An individual's belief in whether an event E occurs can be specified by asking an individual to choose between betting on event E occurring or $A(f)$ occurring. For some frequencies f, the individual will prefer E to $E(f)$ and for others, they will prefer $A(f)$ to E. The point at which the individual is indifferent is called the individual's `subjective probability' for event E.

While conjoint is focused on assessing utility functions for known outcomes, the decisions which will be informed by analysis typically are gambles which do not have guaranteed outcomes. As a result, it becomes important to extend the concept of utility to gambles with uncertain outcomes. To construct these utilities, define an experiment where M is some best possible outcome and m is a worst possible outcome. Consider a gamble which leads to M with frequency f and m otherwise. Again consider a carefully designed laboratory environment where the individual must decide between the consequence and the gamble. Then there will be some maximum value of f for which the individual still prefers the consequence to the gamble. This maximum value measures the individual's preference in the consequence.

In gathering data, it is usually important to have some measure of the confidence which is placed on each of the various data points. To translate this notion of confidence into something tangible, consider two individuals, both whose measure of belief in event E is described by the subjective probability p. Consider a carefully designed laboratory experiment in which each individual observes one success in one trial. Each individual's new belief in the event is then measured. Suppose the resulting value for both individuals is U. A parallel experiment is then run in which the individual's belief in event E is measured after the individual, instead of observing success in the one trial, observes a failure. Let L be the measure of belief which both individuals now have in the event. Now suppose that one of the individual's original assessment of p is based only on observing n trials. *(More precisely, we assume that the individual had a non-informative prior over p and then updated that based on the information in n trials.)* Then it can be shown that $n=1/(U-L)$. Suppose that the other individual's beliefs are based on soft data. Then for analytical purposes, it still is legitimate to use $1/(U-L)$ as a measure of confidence in p.

These examples assumed a carefully designed laboratory experiment. Just as a physical experiment presumes that the physical environment has been prepared to eliminate contaminating influences, so these laboratory experiments must be designed to eliminate contaminating influences like ambiguity, reference point effects, etc.

OBJECTIVE 2. IDENTIFY MEANS OF DATA COLLECTION & ACQUISITION

The focus of this stage is on identifying which kinds of data collection will have the most favorable impact on the quality of the actions and recommendations supported by the analysis. An especially useful tool for doing this analysis is the decision tree. (While the decision tree as applied to uncertainty was formalized in the mid-twentieth century, it can be argued that the Pythagorean Y might have been the first decision tree.) Consider the following very simple decision tree where there are two choices : continue the present course or make a specific change. If a change is made, the outcome of the change could be favorable or unfavorable. We can write this decision tree in outline form as

1. Continue present course

    a. Get an average outcome

    b. Implement a change

2. Get a good outcome

    a. Get a poor outcome

There are two possible outcomes of making the change. If the chance of getting a good outcome is high enough, then it will be better to implement the change. Otherwise implementing the change will be unwise. For example, suppose that we attach a probability p to getting a good outcome if we make a change. Suppose we believe that U is the value (utility) of making a change with the good outcome, L is the value (utility) of making a change if the poor outcome occurs and u is the value (utility) of continuing the present source. Then we will only make a change if

$$p\,U + (1 - p)\,L \; > u.$$

Suppose we find that the best decision (i.e., the one with highest utility to the customer) is to continue the present course. Then we will get utility score u.

But instead of simply making a decision, we could have chosen to gather data and then make our decision based on the results of the data gathering exercise. If we chose to gather data, then our decision tree becomes

1. Gather data and get favorable information

    a. Continue present course

        i. Get an average outcome

    b. Implement a change

        i. Get a good outcome

        ii. Get a poor outcome

2. Gather data and get unfavorable information

    a. Continue present course

        i. Get an average outcome

    b. Implement a change

        i. Get a good outcome

        ii. Get a poor outcome

Now suppose we gather data and get favorable information. This increases the probability of getting a good outcome given we implement a change. Suppose the change in probability is not enough to justify implementing the change. So our conditional decision is *if we get favorable information, we continue with the present course.* Now suppose that instead of getting favorable information, our data gathering led us to collect unfavorable information. This lowers the probability

of getting a good outcome given we implement a change. As a result, our other conditional decision is *if we get unfavorable information, continue with the present course.* Thus our two conditional decisions are *if we get favorable information, we continue with the present course; if we get unfavorable information, continue with the present course.* Hence regardless of the outcome of the information, we continue our present course. This simple example demonstrates an important principle: *Before collecting the information, think about everything you might discovery from collecting the information. If none of these discoveries would lead you to change your decision, then do not collect the information.* Of course, sometimes people collect information—even though they know what decision they will make—in order to defend themselves against criticisms from others. And sometimes people collect information to postpone making the decision.

When would information be valuable? Suppose that the favorable information led to a substantial change in the probability of getting a good outcome. Suppose that this change in probability was enough to justify implementing the change. Then our two conditional decisions would be *if we get favorable information, implement a change; if we get unfavorable information, continue with the present course.* We can assign a value (or utility) to these two conditional decisions. Let u^* be the utility of implementing a change, given that we get favorable information. Let u be the utility of continuing the present course, given that we can unfavorable information. Let q be the probability of our getting favorable information if we collect data. Then the utility if we decide to gather data will be

$$q \, u^* \, + (1 - q) \, u.$$

Since the utility if we did not gather data was u, this tells us that our overall utility has increase from u to $q \, u^* + (1 - q) \, u$. Since $u^* > u$, collecting the information can only improve our utility. This demonstrates a well-established principle: *the value of information is non-negative,* i.e., it can never make you worse off if you behave rationally.

But in reality, there is a cost to collecting this information. Suppose that paying this cost would reduce our utility by some factor d. Thus our utility if we collect information is

$$d \, (q \, u^* + (1 - q) u) \,,$$

while it is u if we do not collect information. So if we knew q and d, the decision on whether to buy information would depend on $(u^* - u)/u$.

What determines u^*? Before making a decision, the chance of getting a good outcome after making a decision was p. Suppose that if we get favorable information, this probability changes to p^* while if we get unfavorable information,

it changes to p**. Then if q is the chance of getting favorable information, the rules of probability require that p=qp*+(1-q)p**. Thus while the utility of making the change was originally

$$p \, U + (1 - p)L,$$

the utility now changes—given a favorable outcomes—to u* where

$$u*=p* \, U + (1 - p*) \, L \quad = p* \, (U - L) + L.$$

So the critical value u* depends upon p* and, in particular, on how much p* differs from p.

The degree to which the new information can change the value of p depends upon the confidence in the original value of p as well as in the impact of the data. One key question is *if the new information tells us something unexpected (i.e. , the favorable outcome), how much will our initial beliefs change?* But given that they do change, we need to know what the potential payoff might be. In this example, H was the maximum payoff if we knew for certain that there would be a good outcome. If the potential payoff, H, were small, then gathering more information would also be pointless.

The final consideration is cost. Since analysts often collect information from the client's subject matter experts, it is important to treat the time of these subject matter experts as precious. If they feel their time is being wasted, then they will complain to the client who will eventually begin to wonder about the value of doing your analysis. There are many cases in which an organization chooses a flawed heuristic over a more sophisticated procedure just because the flawed heuristic seems to require less painful information collection.

There are also privacy issues. Invasion of privacy can lead to a loss of customer good will and, in some cases, legal repercussions. And if we are gathering information that is potentially proprietary intellectual property issues become paramount. The fact that information technology has made it easier to collect information does not mean that information collection is costless.

Once you identify the variables on which you should collect data, the next step is collecting that data. Data collection is analogous to asking certain subjects certain close-ended questions under certain circumstances. Hence there are five steps involved in data collection:

1. Determining how to identify subjects (the sample design)

2. Determining how many subjects to identify (the sampling plan)

3. Determining the questions to be asked

4. Determining the possible answers to the question (the granularity of the experiment)

5. Determining a control group

SAMPLE DESIGN

The population of subjects that could be recruited should be identified. It is common to require random sampling, i.e., to conduct sampling so as to give each subject an equal chance of being part of the sample. This reflects the fact that convenience sampling, e.g., asking those subjects that happen to be easy to identify, has been shown to lead to significant biases. But if the event of interest is highly unlikely, it may be advantageous to bias the sampling toward sampling those individuals most likely to have experienced the event of interest. The analysis will, however, have to take into account this systematic deviation, called stratified random sampling, from conventional random sampling.

Typically each subject has different characteristics (or covariates). To determine how these covariates affect the results of the experimenter, it is tempting—but inefficient—to change one factor at a time and record the change in response from the factor as the impact of that factor. Design of experiments has been shown to be a much more efficient way of assessing the impact of changing factors. This typically involves changing several factors simultaneously. If a full factorial design is used, it is possible to identify the impact of each factor as well as the impact of all possible two-way, three-way, etc. interactions between factors. When it is not necessary to know these higher-order interactions, the less time-consuming fractional factorial designs are used.

It is common to use response surface modeling (and especially regression) to specify the value of interest as a function of the covariates. The independent variable reflects the covariates and are commonly represented using dummy variables for categorical and interval data. When the variable is ratio scale, Box-Cox Transformations are often used to achieve normality. When the dependent variable is categorical, the regression model is typically logistic. When the dependent variable is ordinal, the regression model is typically ordered logit. When the dependent variable is ratio, standard regression is often used. If Y is the dependent variable and $X1,\ldots Xn$ represent the independent (or explanatory variables), then the typical regression model has the form $y=E[Y] + e$ where e is a normally distributed error term and E[Y], the expected value of Y is some parameterized function of $(X1,\ldots Xn)$. In the interests of making this function linear, it is common to write

$$E[Y] = g(a1\ X1' + \dots an\ Xn')$$

where g is a `link' function and X1',…Xn' are monotonic transformations of X1,… Xn. This becomes a generalized linear model if we generalized the error term to be a member of the exponential family of distributions (which includes the normal distribution, the exponential distribution and a remarkably large number of other distributions.)

Because time is often an important dimension, there are a separate body of time-series methods when observations are collected over time. Time-series analysis typically corrects for seasonal patterns (e.g., unusually high sales during holiday seasons) and provides a natural way of identifying trends.

SAMPLING PLAN

How many individuals should be sampled? This is typically determined by the existing amount of uncertainty in the quantity of interest, how much that uncertainty needs to be reduced to facilitate the making of a decision and the degree to which an individual's responses is contaminated with random error. A simple rule of thumb is that quadrupling the number of individuals sampled reduces the uncertainty by half. While uncertainty is commonly measured by the standard deviation, there are situations in which the standard deviation does not exist and the difference between the third and first fractile of the uncertainty distribution is more appropriate.

If we are willing to describe our uncertainty using the previously mentioned exponential family of distributions, the rule for updating uncertainties based on sample information has a very simple form. If our uncertainty is described by an exponential family distribution, it will have two parameters. (In some cases, the parameters may be vectors.) The data is described by the number of observations and the sum of a score for each individual observation. This score for each individual observation will depend upon the exponential family distribution being assumed (If the data consists of coin flips, the score might be one for successes and zero otherwise.) Based on this observed data, the original distribution of the uncertainty is updated. The updated distribution will have the same form as the original distribution with two changes. The first parameter is increased by the summed score while the second parameter is increased by the number of observations. In effect, the original uncertainty about the variable—which reflects soft data—can be treated as if it were generated by a hypothetical set of observations. Pooling this hypothetical data with the actual data then generates a new set of hypothetical data.

DETERMINING THE QUESTIONS TO BE ASKED

A key issue in designing the experiment is determined the nature of the variable being assessed. Is the variable categorical (e.g., values of the variable are blue, red, white) where there is no natural ordering between the values of the variable? When we have categorical scales, the data can be summarized by the proportion of observations which assumed each of the possible values of the categorical variables (e.g., the proportion of blue responses, red responses, etc.)

One can ask YES/NO questions or multiple-choice questions for nominal scales. One extension (likert-type questions) asks subjects to indicate whether they fully agree, partially agree, are neutral, partially disagree or fully disagree with the statement.)

Alternatively the variable might be ordinal (e.g., short, medium, tall) where there is a natural ordering between the values of the variables. When we have ordinal sales, it is possible to define the normalized quantity for each response x by the fraction of responses less than or equal to x (e.g., the fraction of people who are either short or medium.)

A second approach, semantic differential, has the form: `what is your experience navigating our web-site' with answers like `very hard, somewhat hard, OKAY, somewhat easy, very easy' where the two ends of the scale represented opposites. In this case, the response is ordinal.

In both Likert and semantic differential scales, the response scales may be improved by providing concrete examples of what would have to be true for a `fully agree'  or a `fully disagree' response to be true.

A third approach, rank-order, asks individuals to rate various factors in order of importance.

Alternatively the variable might be interval (e.g., thirty degrees centigrade, forty degrees centigrade, fifty degrees centigrade) where the differences between values (e.g., forty degrees minus thirty degrees) are meaningful. Note that when we have interval scales, it is possible to define a normalized quantity for each response x by subtracting the lowest possible value from x and dividing the result by the difference between the highest and lowest value.

A fourth approach is the simple multiple-choice question.

It is important to remember that subjects will often answer a question even when they have no idea about what question is being asked or about what their answer means. (For example, individuals will generally answer the question which is more important diamonds or water even though the answer clearly depends upon

whether the individual feels that the choice is between having no water at all for a week (and dying of dehydration) or simply go without an added glass of water for an hour or two.) Questions must be designed with care.

DETERMINING A CONTROL GROUP

Measurements are typically only meaningful if there is reference to some kind of underlying standard. Thus in extensive measurement, there is some base unit of measure. The score of an item is the number of multiples this basic unit required to create an object that is comparable to the item of interest. By for many non-physical cases, there is no meaningful unit of measurement. In these cases, one creates a benchmark group of units, some smaller than the item of interest and some larger. The score of an item is the proportion of items in a benchmark group which the item outranks. When the item is an uncertain quantity, the score of an item is the probability of the item outranking a randomly chosen item from the benchmark group. This benchmark group is commonly referred to as a control with the item's score being called its effect size.

While some data needs to be created and collected, some data already exists. The purpose of extraction is to collect all this data from the many sources in which it appears so that it can eventually be loaded into a common database. In extracting this data, it is critical to know the data source from which each data element was taken, *i.e., the data must be traceable to its source.* If the results of an analysis depend critically on the data element, then understanding the validity of this data element becomes critical. In addition, if there is some change in the clients for the analysis, it will be important to transition the database to reflect the data sources which these new clients consider important. This requirement is called traceability and typically requires careful documentation.

Another increasingly important issue in using existing data sources is privacy. It is now fairly easy to get personal information on customers by buying such information from vendors. But the customers who provided this information often had an expectation that the information was to be used for a specific purpose, e.g., for enabling them to buy a product over the internet. When these customers discover that their information is being used for another purpose, some customers feel that their privacy is being violated. On top of the privacy issue are intellectual property issues. Even though it may be easy to access the information, there may be copyright or other issues which limit your ability to use it without permission or without compensating the owner of the information.

Data cleaning, while often the least glamorous phase of analysis, is often the most necessary. This is especially the case with pre-existing databases. Because pre-existing databases were collected for other purposes, the quality of the data will be driven by what was important in the original use of this data and hence need not satisfy the quality requirements for the analysis at hand. For example, vendors often have to fill in various forms in order to get reimbursed for their services. Sometimes third parties successfully get their own questions added to these surveys. But both vendor and buyer are primarily interested in the fields which determine how much the vendor gets compensated for their services. As a result, these decision-relevant fields get scrutinized carefully and the rest do not.

There are many other reasons why survey quality may be deficient:

1. Individuals asked to fill out a lengthy survey will get fatigued and simply put in default values so that they can finish the survey. If there are five possible answers to a survey, they may also check a neutral response. Or in a survey of satisfaction, they may either indicate that they are satisfied with everything or satisfied with nothing.

2. Individuals may also be offended by questions about their age, income, marital status, ethnicity and—if the survey forces them to fill in an answer—will often deliberately fill in a false answer. This is especially true given increasing concerns about privacy

3. Biases can often arise because most people, when asked to fill out a survey, simply refuse. Those who did fill out the survey are often people with more leisure time or with more emotional commitment to the organization asking that the survey be filled out.

As a result, data cleaning includes:

1. Identifying the range of valid responses for each question and labeling the data field

2. Identifying invalid data responses (e.g., where letters are used where numbers are required)

3. Identifying inconsistent data encodings (e.g., different abbreviations might be used for state)

4. Identifying suspicious data responses (e.g., when physically questionable numbers are put in for a response) Are there outliers that don't seem to make sense?

5. Identify suspicious distribution of values (e.g., when one finds that 99% of the respondents in a survey of poor neighborhoods have incomes of more than a million dollars.) Descriptive statistics can be very helpful in identifying suspicious distributions. For example, histograms specify the frequency with which various data response are used. Box and Whisker charts as well as stem and leaf plots provide compact descriptions of the variation in the data within a field and help identify outliers. Scatterplots show how the value of one set of variables depends on another. Summary statistics like the mean, median, upper and lower fractiles can also be useful

6. Identify suspicious interrelationships between fields. We first identify whether there is any correlation between data fields—possibly using factor analysis or principle component analysis. The creator of the database may have created a new variable—by combining existing fields—which was useful for their analysis but is no longer useful for your analysis

So a key part of data cleaning is determining whether the data makes sense. It also involves handling null or missing values. There are several possible solutions:

1. **Deletion:** Dropping the observation containing the missing value

2. **Deletion when necessary:** Not using the observation in analyses requiring a valid response for the missing item. This approach means that one might have a sample of 1000 people for one kind of analysis and a sample of 950 people for a second kind of analysis.

3. **Imputing a value:** In other words, we use regression to attempt to predict what the answer to this question would have been—based on the answers the subject gave to other questions.

4. **Randomly imputing a value:** The problem with imputing a value is that it pretends that we do know the value that the subject filled it for this question. Thus understates our uncertainty about the value (and thus overstates our sample size) which can lead to biases in the analysis. Random imputation in theory reruns the analysis for all possible responses the subject might have given to the question, weighted by the regression-based probability of the subjective giving that response. Efficient algorithms have been developed for doing multiple imputation.

It is important to determine whether important observations (e.g., observations from a specific group of sub-users) is missing.

A field should be created with the data of each observation (a date stamp.) A field should also be created identifying the data source from which this information is collection. This field will be important in the next step where information from different data sources is combined into a single database

While the individual responses come from different data sources, they need to be placed into a common database (which typically is organized into rows representing observations and columns representing observed characteristics of that observation). This requires that all of the data be summarized at a common level of granularity.

For example, we might have 1000 observations of one product, 5000 observations of a product and its location and 3 observations of a product, its option content and its location. If details about a product's location are not relevant for the analysis, then we can sum up our observations so that all data is at this less granular level. In other cases, we need to go to this more granular level. If we simply dropped all the observations that did not have this information, there could be insufficient sample size to support a meaningful analysis. Alternatively we may rewrite all of our 9000 records at this more granular level with fields for the product, its location and its option content. We now must treat many of our records as if they had missing values for location and option content.

Sometimes data is aggregated in different ways. Thus some information on vehicles is stored with certain vehicles being called two-door Chevrolets. Other information is stored with certain vehicles called Chevy Cruzes. Still other information is stored as General Motors compacts. In this case, there is no a single categorization that is more granular than any other categorization. As a result, we may simply need to define a record which has enough fields to contain the information from each of these observations. Thus there might be a field indicating whether the vehicle was two-door or not, a field for the vehicle's model name (Cruz), a field indicating the vehicle's body-type (compact) and a field reflecting the vehicle's division and manufacturer.

In some cases, the model may require information on a variable which is not in the database but can be computed from items in the database. This may require the creation of a new field in the database for this derived variable.

In some cases, a single observation may reflect the responses of 10,000 people while another observation may reflect the responses of 100 people. As opposed to creating a database with 10,100 rows for these two observations, it may be useful to introduce a weighting field that identifies the number of respondents associated with the observation.

Because different datasets are typically generated with different data architectures and different programming languages, these languages may use different standards for encoding information. Thus missing values can be represented by spaces, the words NA, the words Not/Available, etc.

Some decisions may be required in how to handle textual fields. This could be handled by creating numeric columns describing the textual field and—without deleting the textual field—using the columns to classify the field. For example, the textual field might contain verbatim user expressions of satisfaction. A column might be created which expresses the encoder's interpretation of that field as expressing satisfaction, dissatisfaction or neutrality.

Before loading the database, it is useful to assess whether certain fields have the same value across all datasets. If this is the case, then it may be worth deleting those fields.

The data is then loaded into the common database. Information is typically normalized so that any given item of information only occurs in the database exactly once. This is the place to do some final checks on the quality of the data:

1. Completeness: Are all the fields of the data complete?

2. Correctness: Is the data accurate?

3. Consistency: Is the data provided under a given field and for a given concept consistent with the definition of that field and concept?

4. Currency: Is the data obsolete?

5. Collaborative: Is the data based on one opinion or on a consensus of experts in the relative area?

6. Confidential: Is the data secure from unauthorized use by individuals other than the decision maker?

7. Clarity: Is the data legible and comprehensible?

8. Common Format: Is the data in a format easily used in the application for which it is intended?

9. Convenient: Can the data be conveniently and quickly accessed by the intended user in a time-frame that allows for it to be effectively used?

10. Cost-effective: Is the cost of collecting and using the data commensurate with its value?

The term data warehouse is generally used to describe:

1. A staging area, i.e, the operational data sets from which the information is extracted

2. Data integration which is the centralized source where the data is conveniently stored

3. Access layers, i.e., multiple OLAP (on-line analytical processing) data marts which store the data in a form which will be easy for the analysis to retrieve

The data mart is organized along a single point of view (e.g., time, product type, geography) for efficient data retrieval. It allows analysts to

1. slice data, i.e., filtering data by picking a specific subset of the data-cube and choosing a single value for one of its dimensions;

2. dice data, i.e., grouping data by picking specific values for multiple dimensions;

3. drill-down/up, i.e., allow the user to navigate from the most summarized (high-level) to the most detailed (drill-down);

4. roll-up, i.e., summarize the data along a dimension (e.g., computing totals or using some other formula);

5. pivot, i.e., interchange rows and columns (`rotate the cube').

Fact tables are used to record measurements or metrics for specific events at a fairly granular level of detail. Transaction fact details record facts about specific events (like sales events), snapshot fact tables record facts at a given point in time (like account details at month end) and accumulating snapshot tables record aggregate facts at a given point in time. Dimension tables have a smaller number of records compared to fact tables although each record may have a very large number of attributes. Dimension table includes time dimension tables, geography dimension table, product dimension table, employee dimension table, and range dimension tables.

Each dimension is typically ranged into hierarchies, e.g., the geography dimension might be arranged in stores, cities, states and countries. These hierarchies are often dynamic, e.g., a firm may redraw its organizational boundaries. In the star schema, there is often a single fact table with many dimensional tables surrounding it.

The leads to a data mart which will service the analysts in an efficient matter. However the data warehouse and data marts are not finished until they are documented in a way that makes them usable by external parties. While it is tempting to assume

that the modeler will know what the variables mean, the reality is that there will often be requests to revisit the data months or years after the analysis is done. These requests may come from the client or they may come from peer reviewers interested in replicating your work. In this case, failure to document your data fields as well as the sources of the data can be very costly.

OBJECTIVE 4. IDENTIFY WAYS OF DISCOVERING RELATIONSHIPS IN THE DATA

The many ways of understanding data can be organized into nine steps. The following list from Booz-Allen-Hamilton's The Field Guide to Data Science describes some of the techniques which can be useful in implementing each of these steps:

1. Filtering

    a. Filtering can involve using relational algebra projection and selection to add or remove data based on its value.

    b. Filtering usually involves outlier removal, exponential smoothing and the use of either Gaussian or median filters.

2. Filling in missing data with imputation:

    a. If other observations in the dataset can be used, then values for missing data can be generated using random sampling or Monte Carlo Markov Chain methods.

    b. To avoid using other observations, imputation can be done using the mean, regression models or statistical distributions based on existing observations.

3. Reducing the number of dimensions in the data:

    a. Principle component analysis or factor analysis can help determine whether there is correlation across different dimensions in the data.

    b. For unstructured text data, term frequency-inverse document frequency identifies the importance of a word in some document in a collection by comparing the frequency with which the word appears in the document to the frequency with which the word appears in the collection to which the document belongs.

    c. When data has a variable number of features, feature hashing is an efficient method for creating a fixed number of features which form the indices of an array.

d.   Sensitivity analysis and wrapper methods are typically essential when you don't know which features of your data are important. Wrapper methods, unlike sensitivity analysis, typically involving identifying a set of features on a small sample and then testing that set on a holdout sample.

e.   Finally self-organizing maps and Bayes nets are helpful in understanding the probability distribution of the data.

4.   Extracting features

a.   Duplicate data elements must be corrected with de-duplication methods.

b.   Normalization is required to ensure your data stays within common ranges. This prevents the scales in which data was collected from obscuring the interpretation and analysis of that data.

c.   Format conversion is typically required when data is in binary format.

d.   Fast Fourier Transforms and Discrete wavelet transforms are used for frequency data.

e.   Coordinate transformations are used for geometric data defined over Euclidean.

5.   Collecting and summarizing data

a.   Basic statistics (raw counts, means, medians, standard deviations, ranges) are helpful in summarizing data.

b.   Box plots, scatter plots, box and whisker plots provide compact representations of how data is distributed. But when the data can be reasonably described by parametric distributions, distribution fitting are even more efficient ways of summarizing data.

c.   `Baseball card' aggregation is an effective way of summarizing all the information available on an entity.

6.   Adding new information to the data

a.   Annotation is recommended for tracking source information and other user-defined parameters.

b.   Relational algebra rename and feature addition (e.g., geography, technology, weather) can be helpful in processing certain data fields together or in using one field to compute the value of another.

7. Segmenting the data to find natural groupings

   a. Connectivity-Based methods: Hierarchical clustering generates an ordered set of clusters with variable precision.

   b. Centroid–Based methods: When the number of clusters is known, k-means is a popular technique. When the number is unknown, x-means is a useful extension of k-means that both creates clusters and searches for the optimal number of clusters. Canopy clustering is an alternate way of enhancing k-means when the number of clusters is unknown.

   c. Distribution-based methods: Gaussian mixture models, which typically used the expectation-maximization (EM) algorithm, are appropriate if you want any data element's membership in a segment to be `soft.'

   d. Density-based methods: For non-elliptical clusters, fractal and DB scan are useful.

   e. Graph-Based methods: Such methods, often based on constructing cliques and semi-cliques, are useful when you only have knowledge of how one item is connected to another.

   f. For text data, topic modeling allows for segmentation of the data

8. Determining which variables are important

   a. When the structure of the data is unknown, tree-based methods are helpful.

   b. If statistical measures of importance are needed, generalized linear models are appropriate. But if statistical measures of importance are not needed, regression with shrinkage (e.g., LASSO, elastic net) and stepwise regression may be preferable.

9. Classifying data into existing groups

   a. If you are unsure of feature importance, neutral nets and random forests are helpful. But if you require a highly transparent model, decision trees (e.g., CART, CHAID) can be preferable.

   b. When the number of data dimensions is less than twenty, k nearest neighbor methods often work. But if you have a large dataset with an unknown classification signal, naïve Bayes may be preferable.

c. Hidden Markov models are useful in estimating an unobservable state based on observable values.

OBJECTIVE 5. DETERMINE THE DOCUMENTATION & REPORTING OF FINDINGS

Learning Objectives 5 and 6 go together. Raw data and relationships, while interesting to analysts, will not hold the attention of your business stakeholders for long. You will need to tie your findings to the analytics problem and from there to the business problem. Going back to the example in the business problem domain of the manufacturing plant, a key relationship unearthed by your data analysis might be an inverse proportionality of WIP to queue and wait time. This would need to be communicated clearly to your stakeholders along with a recommendation to reduce the rate at which material is being released to the floor to enable faster delivery. Another key relationship might be a workcenter's scrap rate being higher than its normal or "should fail" rate.

OBJECTIVE 6. USE DATA ANALYSIS RESULTS TO REFINE BUSINESS & ANALYTICS PROBLEM STATEMENTS

Having solid data and relationships allows the first true refinement of your analytics and business problem, as you now have the ability to go beyond anecdotes of the situation and describe the situation with some level of mathematical rigor. You may find at this point that the true constraint of the system isn't what you thought it was, and that therefore the analytics problem needs to be reframed around that newly surfaced constraint. Or you may find that the business problem itself missed a key facet (interrelationships between customers and purchases, a time-series effect in the data, or anything else) that needs to be included prior to continuing. Once in a while, you actually do get the business problem and the analytics problem right the first time and you can proceed to selecting your methodology and creating your model.

SUMMARY

It is no accident that the CAP exam weights data the most heavily of the seven domains. Without proper data gathering, cleaning, transformation, and loading, all you have are nice anecdotes. With reliable data sorted usefully, you can actually solve your problem in a meaningful way.

Booz-Allen-Hamilton, 2013, The Field Guide to Data Science, http://www.boozallen.com/media/file/The-Field-Guide-to-Data-Science.pdf.

Hubbard DW (2010) *How to Measure Anything: Finding the Value of "Intangibles" in Business*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Hillier F, Hillier M (2010) *Introduction to Management Science: A Modeling and Case Study Approach*, 4th ed. (McGraw-Hill Higher Education, New York).

Vose D (2008) *Risk Analysis: A Quantitative Guide*, 3rd ed. (John Wiley & Sons, Chichester, UK).

WHAT WILL YOU LEARN IN THIS CHAPTER?

In this chapter, you will learn about examples of various methods available to analytics professionals and how to go about choosing some of them over others for a specific task. This chapter does not intend to offer an exhaustive list of such methods; instead it is rather illustrative to convey the process of selection from some methods. There are a myriad of analytics methodologies in the literature that are available from which a modeler can select. Later in this chapter, we will list typically used analytics methodologies, their characteristics, classifications, and when to use them.

Selection (of methods) process is best informed by the problem framing, prior experience and depth of knowledge of the analytics professional with the methods available, the problem at hand, etc. However, it is very possible that the problem at hand is rather new or not framed completely. In such situations, selecting the best methodology to solve a problem could be iterative in nature since a methodology may prove to be ineffective so other methodologies may need to be tried as well. It is often the case that an analyst (the modeler) is not given enough time to explore many options. Therefore, utilizing experience and knowledge will certainly help to improve the chance of selecting an effective method in a timely manner.

Learning Objectives:

1. Identify available problem solving approaches (methods)

2. Select software tools

3. Test approaches (methods)*

4. Select approaches (methods)*

OBJECTIVE 1. IDENTIFY AVAILABLE PROBLEM SOLVING APPROACHES

Almost all analytical models can be classified into one of three categories: descriptive, predictive, and prescriptive. These three categories of models do as their names imply.

Prescriptive methodologies offer solutions that provide specific quantifiable answers that can be implemented to solve a problem. For example, a linear programming model is a prescriptive model.

A prescriptive model answers the question "What is the best action or outcome?" See Exhibit 1. The key factor here is to provide new ways to improve certain types of performance as agreed upon with the customer and documented in the business and analytics problem statements. Some examples of prescriptive techniques are:

- Optimization
    - Linear programming
    - Integer programming
    - Nonlinear programming
    - Mixed integer programming
    - Network optimization
    - Dynamic programming
    - Metaheuristics
- Simulation-Optimization
- Stochastic Optimization

Predictive methodologies include any forecasting models such as time-series models, for example, moving averages, auto-regression models, etc. These models mostly make a forecast for the future to answer the question "What could happen?" See Exhibit 1. The key factor here is to predict future trends and possibilities. Some examples of predictive techniques are:

- Simulation
    - Discrete event
    - Monte Carlo
    - Agent-based modeling
- Regression
    - Logistic
    - Linear
    - Step-wise

- Statistical Inferences

  - Confidence intervals

  - Hypothesis testing

  - Analysis of variance

  - Design of experiments

- Classification

- Clustering

- Artificial Intelligence (AI)

- Game Theory

Descriptive methodologies are a collection of models that help describe the problem situation for further analysis. It can be based on descriptive statistics that are conveyed through (1) charts and graphs such as histograms, scatter plots, etc., and (2) numerical presentations such mean, median, mode, variance, standard deviations of distributions of data, and cross tabulations. They answer the question "What happened?" See Figure 5. The key factor here is the use of historical data.

PRESCRIPTIVE ANALYTICS	**What is the best action/ outcome?**	• Evaluates and determines *new* ways to operate • Targets business objectives • Balances all constraints	**Very High**
PREDICTIVE ANALYTICS	**What could happen?**	• Predicts *future* probabilities and trends • Finds relationships in data not readily apparent with traditional analysis	**Data Needs Benefits**
DESCRIPTIVE ANALYTICS	**What happened?**	• Prepares and analyzes *historical* data • Identifies patterns from samples for reporting of trends	**Medium**

Figure 5: INFORMS CAP® Methodology Classification

The following are the primary factors that an analyst generally considers to select an appropriate methodology:

1. Time–Typically modelers work under tight timelines. They are faced with the challenge of choosing the right methodology and quickly running their modeling and analysis to answer the business needs.

2. Accuracy of the model needed – Some models (and their level of aggregation) influence the accuracy of the results. This is closely related to the quality and readiness of data in addition to the level of accuracy needed for the model as requested by customer. If available data are not accurate, using a very accurate model may be a waste of time. For instance, a modeler is advised not to seek optimal solutions where he or she has more noise in the data than the signal of the optimal answer.

3. Relevance of the methodology and scope of project–Refer to INFORMS CAP® methodology classification in Exhibit 1. Often the business problem statement and the analytics problem statement (domains I and II in INFORMS CAP®) can be suggestive of the relevance of the chosen methodology classification. If a customer is seeking to understand the most influential variables to impact other ones, then descriptive models would be the most relevant. Predictive methodologies are the most relevant when the purpose of a project is to forecast certain behavior. Prescriptive, on the other hand, are most helpful when there is a need to pinpoint certain decisions to the level of quantifying the variables that enhance the performance under study.

4. Accuracy of the data–Sometimes data level of aggregation and their accuracy dictate the pool of methodologies to be used. Very accurate methodologies may not be the best way to go when the accuracy of data is low.

5. Data availability and readiness–Sometimes data availability plays a large role in the selection process.

6. Staff and resource availability–The nail-and-hammer metaphor applies very well here. A staff of statisticians may focus on the use of statistics in their services; and a staff of operations research professionals may focus on the field of simulation and optimization in their services. A collaborative and diverse team is the best way to get around such challenges.

7. Methodology popularity–Some methodologies are well known in the market and are focused on and sometimes demanded by the customer to be used. It is the modeler's responsibility to use the best approach rather than going with what is popular and demanded.

Some commonly used methods/models

There are countless methodologies in the literature from which a modeler can choose to solve a problem. Here are a few types of analytics methodologies that are commonly used:

- Discrete event simulation. A simulation methodology that has the following characteristics:

    ▪ Often used to understand bottleneck(s) in systems

    ▪ Handles cases that cannot be handled by queuing theory

    ▪ Often used for multistage processes modeling with variations in their arrivals and service time and utilizing shared resources to perform multiple operations

- Queuing model. Designed to identify the most efficient pathway to solution; i.e., at a bank it might identify the number of tellers needed to satisfy customers in a particular time frame such as no more than 10 minutes waiting in a queue.

- Monte Carlo simulation. Characteristics include:

    ▪ Queuing modeling is not needed

    ▪ Used primarily to estimate dependent variable randomness out of a set of independent variable randomness. This is especially necessary when distributions of the input variables are not necessarily normally distributed and the relationship to estimate the dependent variable is not simple (e.g. additive)

- Agent-based modeling (ABM). A system modeled (simulated) as a collection of autonomous decision-making entities called agents that are used to discover emergent behavior that is hard to predict without simulating it.

- System dynamics (SD). A simulation approach used to understanding the interactions of a complex system over time.

- Game theory. Study of strategic decision-making processes through competition and collaboration.

- Probabilities. The likelihood of a particular event occurring expressed as a percentage to make decisions under chosen risk or tolerance. Bayesian and conditional probabilities are widely used in analytics.

- Economic analysis. Evaluation often used to guide the optimal allocation of scarce resources:

  - IRR (internal rate of return)–discounted rate used in capital budgeting to compare returns on investment opportunities

  - NPV (net present value)–difference between present value of incomeversus outgo

  - FV (future value)–value of a future event or item based on current value that is adjusted by some standard amount, i.e., expected inflation

  - Payback period–period of time after which an expenditure is fully amortized and income begins to accrue in excess of expenses

- Regression. A class of statistical methodologies used to map dependent variables with independent variables and understand the significance between the variables and their correlations with one another.

  - Linear regression–compares the relationship between a dependent variable and one or more explanatory variables. The variables here are linear; however nonlinear functions can be explored here by scaling input data.

  - Stepwise regression–method of model building that successively adds or deletes variables based on performance.

  - Logistic regression–may also be called logit analysis, is a regression analysis often used to predict the outcome of categorical variables.

- Statistical inferences:

  - Confidence intervals

  - Hypothesis testing

- Design of experiments

- Analysis of variance (ANOVA)

- Principal component analysis (PCA). Can be used to reduce data dimensionality.

- Data mining

- Forecasting

- Artificial intelligence:

  - Artificial neural networks

  - Fuzzy logic

  - Expert systems

- Decision trees

- Optimization

  - Linear programming

  - Integer programming

  - Mixed integer programming

  - Combinatorial optimization

  - Nonlinear programming

  - Constraint programming

  - Response surface methodology (RSM)

  - Metaheuristics

  - Greedy heuristics

- Markov chain

- Revenue management (yield management)

Different methodologies target problems at different levels. For example, using value stream mapping requires more aggregate data compared to a discrete-event simulation model. A problem can be solved at different levels of aggregation. A supply chain problem, for example, can be modeled on the product-family level or on the SKU level. Other examples of aggregation are: assemblies versus subassemblies, region versus country versus state versus city level, etc. The lower the level of aggregation, the more accurate and descriptive the model will be of the real-life scenario; however, it will be harder to validate and will certainly be more prone to mistakes. On the other hand, a higher level of aggregation usually provides faster results that are easier to understand. The general rule of thumb is to model at the highest level of aggregation possible that will ensure a satisfactory level of accuracy within the time permitted.

Even if you have all the time in the world, the knowledge of all methodologies, and availability and accuracy of the needed data, it is highly desirable and advisable to run scenarios on the "back of an envelope"–often referred to as quick and dirty (Q-n-D). That approach may provide the high level understanding needed to make a decision quickly to go with certain strategies and/or orient the applied methodology accordingly. In all of these endeavors, it is important to communicate with your stakeholders in a way that they'll understand your approach and its pros and cons. Often, your managers will have much less deep analytics knowledge than you, but they will likely understand the business needs better.

A key point for the CAP® is that the exam is vendor and toolset neutral. The society is looking for understanding of how to apply tools, not certifying people in the use of a particular tool. A good analyst will have a tool chest with several different tools to fit various situations.

Here are software categories from an analytics point of view as we see it:

- Spreadsheet systems

- Statistical systems

- Optimization systems

- Simulation systems

- Business intelligence systems

- Data management systems

  - Structured data

  - Unstructured data

- Data integration systems

- Operating systems such as HADOOP

Figure 6 shows a set of sample software applications that are used in analytics and being compared against different aspects.

SOFTWARE TOOL	VISUALIZATION	OPTIMIZATION	SIMULATION*	DATA MINING	STATISTICAL
EXCEL	HIGH	LOW	LOW	MEDIUM	MEDIUM
ACCESS	LOW	LOW	LOW	MEDIUM	MEDIUM
R	LOW	LOW	LOW	HIGH	HIGH
MATLAB	MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM
FLEXSIM	HIGH	LOW	HIGH	LOW	MEDIUM
PRO MODEL	MEDIUM	LOW	HIGH	LOW	MEDIUM
SAS	MEDIUM	HIGH	MEDIUM	MEDIUM	HIGH
MINITAB	MEDIUM	LOW	LOW	LOW	HIGH
JMP	MEDIUM	HIGH	MEDIUM	MEDIUM	HIGH
CRYSTAL BALL	MEDIUM	LOW	HIGH	LOW	MEDIUM
ANALYTICA	HIGH	HIGH	MEDIUM	LOW	LOW
FRONTLINE	LOW	HIGH	LOW	LOW	LOW
TABLEAU	HIGH	LOW	LOW	MEDIUM	LOW
ANYLOGIC	LOW	LOW	SIMULATION	LOW	LOW

This includes: Monte Carlo, discrete-event, system dynamics, and agent-based simulation. Figure 6. Sample software application characteristics

OBJECTIVE 3. MODEL TESTING APPROACHES

Developed models need to be both verified and validated. The verification step refers to making certain that the model is built the way it was designed and meant to be. The validation step refers to making certain that the model is representing real life to a certain level of accuracy. If the modeler realizes that validation and verification are unachievable, then the modeler needs to consider other approaches.

To help the testing process, it is advisable to divide data into three portions:

- Building – This portion of data is used to estimate the needed parameters such as slopes in case of regression.

- Testing – This portion of data is used to test the model (verify) that was it was modeled as it was designed.

- Validating – This portion of data is used to test that the model behaves closely to the physical behavior being modeled.

As this topic is not covered in the CAP®, suffice it to say that after you test your models, it makes sense to go with the most accurate model that otherwise complies with your time and cost constraints.

SUMMARY

In this chapter, we covered how a class of analytics is chosen and then how a particular methodology as well a supporting environment (software) is selected. There is certainly no one methodology to choose but there are ones more fit than others based on the available time, data, expertise, and relevance.

The following statements summarize the main points covered in this chapter:

- The methodologies used to solve the problems are scientific; however selecting the appropriate one is a combination of art and science.

- A problem may be solved using more than one methodology.

- Choosing the level of detail (aggregation) of the model is important. The general rule to follow is to model the problem at the highest level at the desired level of accuracy.

- Available approaches vary in their accuracy. Choosing an approach shall match with the accuracy of data.

- A world-class organization is the one that employs a diverse team of analytics professionals with knowledge depth and breadth.

FURTHER READING

Big data: The next frontier for innovation, competition, and productivity, a McKinsey & Company report. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

WHAT WILL YOU LEARN IN THIS CHAPTER?

Model building is at the heart of any analytical effort; it is the climax of the analytics problem framing activities. Good models depend on all previous steps: framing the business problem; framing the analytics problem; and acquiring, exploring, and scrubbing the data. Now it is time to develop a model to show key drivers of your outcomes, forecast your targets, determine the best use of resources, etc. Effective model building requires identifying the relevant inputs and selecting the model that performs best on holdout or testing data. The emphasis in this chapter is on statistical methods for predictive modeling.

You have learned to frame the problem, acquire and clean the data, and select a methodology or modeling approach to solve the problem. Now it is time to build the model. This chapter introduces the process of model building in a business analytics context. You will learn to identify model structures appropriate to your analytical objective. Different model structures require different data characteristics. You will also learn about the importance of honest assessment and data splitting to assess models. You will learn to select a champion model from several candidates, and to communicate the key findings to stakeholders.

Learning Objectives

1. Identify and build effective model structures to help solve the business problem

2. Run and evaluate the models

3. Calibrate models and data*

4. Integrate the models*

OBJECTIVE 1. IDENTIFY MODEL STRUCTURES

Once a methodology is determined, it still remains to build the model. Sometimes this means refining the methodology. For example, if you have determined that a predictive data mining model is the right approach to solve your problem, then you need to determine the specific type of model that will perform best in predicting your target. Does the target follow a binomial distribution? Normal? Gamma? Poisson? Is the target censored? Within a class of models appropriate to the target, there can be many specific types of models to choose among. A

logistic regression, a decision tree, and a neural network can all predict in binary target. This is not typically done *a priori*. Instead, you might identify several types of models, fit all of these models to the data, and select a champion.

By this point in the project plan, you should have collected the data or set out a data collection plan. However, much of the work with data occurs alongside the model building process. Different models assume particular data structures. Do you need transactional data? Individual-level data? Household? Do your values have a time component? What kinds of summary statistics should be used to roll up values from lower to higher levels? If transactions are coded in dollars, should a household-level value consist of a sum, average, maximum, or something else? The answers to these questions can depend on the business objective and what you intend to learn from that variable; they can also depend on the class of models you have selected.

Building the models and working the data into an appropriate form require collaboration among the analyst, the data owner, and the subject matter expert. The subject matter expert should have a clear vision for the types of characteristics needed for modeling. Demographics, historical behavior, attitudinal surveys, and other characteristics should be identified and carefully selected by someone who understands the business problem. This was likely addressed while framing the analytics problem statement. The analyst must pay close attention to data quality requirements for modeling. For example, in some models, data should be equally spaced, missing values should be handled, variance stabilizing transformations should be applied, and so on. This is more likely to be addressed at the time of model building than during data acquisition. The data owner needs to know how to bring the characteristics together, from potentially disparate sources, to create the data structure that the analyst requires. Some of this work will be done early in the project, but much of the data cleaning necessarily occurs at the time of model development, as each modeling type has its own data obstacles.

You have determined the model type(s) and gathered the appropriate data. The next step is building and refining the model. Building the model is, for many analysts, the most enjoyable part of an analytical project. Although it can require a great deal of work to define a simulation model, thinking through the relationships and identifying all relevant sources of variability is problem solving at its best. There is invariably a sense of anticipation in preparing to evaluate the results of a predictive model and its assessment on holdout data.

If you have fit several models, then it will be necessary to perform an honest assessment of their performance so that a champion can be selected. This is discussed next. One key consideration in running models is how the models will be used later. For example, a model that will result in scoring activity should have a way to score new observations without refitting the model or estimating

new parameters. Preferably, it should be possible to perform scoring in a real-time production environment where specialized analytical software might not be available.

OBJECTIVE 2. EVALUATE & CALIBRATE MODELS & DATA

Model performance can be evaluated in a number of ways. A model can perform well in describing a phenomenon, predicting a target, or optimizing characteristic settings for a system. The same model is not likely to be best for all three purposes. As such, it is necessary to select a model based on metrics that align with what the model must accomplish.

For example, a predictive model should always be selected using an honest assessment of the model on holdout data. If you have fit a variety of models to your sample data and must select a champion, you will need to define a "goodness" metric appropriate to how the model will be used. If the goal is to correctly classify observations on a binary target, that metric might be misclassification, sensitivity, specificity, or other similar metrics based on correct and incorrect classifications. A model that will be used to select the "top x%" from a sample should be assessed using a metric that evaluates the rank order of predicted values, such as concordance, discordance, ROC/c-statistic, among others.

You might be selecting a champion from a series of models of increasing complexity from one model type (e.g., regression). Alternatively, you might have a selection of models that have been fit to the data (e.g., a regression, decision tree, and a neural network). In either case, selecting a champion follows the same method: honest assessment on a holdout sample of data.

Honest assessment techniques can vary, and might include data splitting, k-fold cross validation, leave-one-out cross validation, etc. What is critical in honest assessment is that the observations used to fit the model and estimate parameters are not the observations that are scored in assessment. The process, conceptually, is relatively simple. Honest assessment with data splitting on a binary target is described as follows.

1. Select a large sample of data for modeling. For a binary target, a good practice is to ensure that you have at least 2000 observations in the smaller of the two target classes.

2. Use a stratified random sampling without replacement to create two data sets with approximately the same proportion of 0 and 1 target levels. The sample need not be equally sized. For example, some analysts prefer a 70/30 split into training and validation sets as this allocates more observations to training and makes parameter estimates more stable.

3. Fit models and estimate parameters using the training data.

4. Using the assessment statistic you considered previously, score observations in the validation data set. If the model uses stop training, pruning, or model selection with stopping rules, then those selections should always be based on the validation data performance.

Selecting the champion model can be as simple as selecting the model with the best performance; alternatively, you might select the champion based on a combination of model performance and interpretability. Some models, such as neural networks, might not be selected (because of their difficulty in interpreting the model), but might be used as a benchmark against which other models are compared.

Not all models are supervised or have data labeled with predefined classes. Some are unsupervised with unknown class labels for the data. Examples of unsupervised techniques commonly used in business analytics include: segmentation through clustering, rule generation through market basket/association analysis, deriving links among nodes through social network analysis, measurement of latent variables through common factor analysis, etc. Unsupervised analyses should also be empirically validated to ensure that your findings reflect more than the idiosyncrasies of the sample. However, the techniques for validating unsupervised analyses are not as straightforward and typically rely on the analyst's best judgment.

OBJECTIVE 3. CALIBRATE MODELS & DATA

Once your champion model is selected, it is time to improve both the model and the data approach to refine the model. This may be something as simple as recognizing that really you need to take into account time series information as well as household transactions per year and reformulating the data structure to take that into account. Or you may find that there is a subsegment of your population that your model really doesn't measure well and you need to create a subsidiary model for that subsegment.

A key concept here is managing the tension between "I need an answer" and "I don't fully trust the model yet." Your business stakeholders chartered this project because they need an answer. Every day that passes makes that need more acute. At the same time, you as the analyst know the strengths and weaknesses of your model in a way that your stakeholders may not appreciate. Negotiating a reasonable level of confidence up front can help with this, along with communicating your plan of how to get from where you are to where you need to be.

OBJECTIVE 4. INTEGRATE THE MODELS

Although this topic is beyond the scope of the CAP® exam, model integration is needed when you are bringing a new model into an existing model environment. Often your model will take outputs from other models, and its output will feed inputs. Documenting your inputs and outputs in an API (application programming interface) like schema will help with that integration.

SUMMARY

The model is the heart of the answer to the business analytics problem. Build it carefully, test it thoroughly, calibrate it properly, but be willing to tear it down and start over again to get a truthful answer to the business problem.

FURTHER READING

Berry MJA, Linoff GS (1999) *Mastering Data Mining: The Art and Science of Customer Relationship Management* (Wiley, New York).

Clemen RT (1997) *Making Hard Decisions: An Introduction to Decision*, 2nd ed. (Duxbury Press, Pacific Grove, CA).

Few S (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, 2nd ed. (Analytics Press, Burlingame, CA).

Hand DJ, Mannila H, Smyth P (2001) *Principles of Data Mining* (MIT Press, Boston).

Hillier FS, Lieberman GJ (2005) *Introduction to Operations Research*, 8th ed. (McGraw-Hill, New York).

Law AM, Kelton DW (2000) *Simulation Modeling and Analysis*, 3rd ed. (McGraw-Hill, New York).

Ross SM (2010) *Introductory Statistics*, 3rd ed. (Academic Press, Burlington, MA).

Siegel E (2013) *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (Wiley, New York).

Tufte ER (2001) *The Visual Display of Quantitative Information*, 2nd ed. (Graphics Press, Cheshire, CT).

THIS PAGE IS DELIBERATELY LEFT BLANK

WHAT WILL YOU LEARN IN THIS CHAPTER?

As described in previous chapters, analytical solutions go through a lifecycle from conception and business justification through data collection, model construction, and deployment. This chapter focuses on the deployment phase. An analytical model must be deployed to be useful to the organization. Deploying the model involves interacting with the business users who will use the model directly or the results of the model (e.g., if the model was the basis of a study report).

An effective deployment requires careful planning so that all staff involved know their roles in the deployment, and are trained so they appropriately use and interpret the results.

This chapter introduces the deployment stage as illustrated by the standard CRISP-DM (cross-industry standard process for data mining) as well as several considerations for deployments within the enterprise. You will learn the CRISP-DM methodology for the deployment stage. You will also learn additional considerations for deploying a real-time analytics model.

The CRISP-DM standard is not the only tool for deployment. For example, Six Sigma's define, measure, analyze, improve, control (DMAIC) methodology includes some of the same concepts, especially in the sections on proposed solution, piloted solution, and sustained solution.

Learning Objectives

1. Perform business validation of the model

2. Deliver report with findings; or

3. Create model, usability, and system requirements for production

4. Deliver production model/system*

5. Support deployment

    a. Learn the industry standard CRISP-DM deployment approach

    b. Learn about other considerations for deployment of real-time models

    c. Understand the considerations needed to plan the deployment of analytics in your organization

Key Concepts/Fundamentals

- The CRISP-DM deployment stage

    ▪ Produce Final Report – Depending on the project, this may be more or less comprehensive. For some one-time projects, this may document the results of the analysis.

    ▪ Review Project – All projects should be reviewed for what went right or wrong, and what should be improved in the future.

OBJECTIVE 1. PERFORM BUSINESS VALIDATION OF THE MODEL

Quite simply this comes down to making sure that your answer is still tied to the original question. It is not uncommon for discrepancies to creep in to the analysis as the problem is framed and communicated. It is also not uncommon for the business context to have changed since the project started, invalidating key assumptions. While the above items must be taken into account, be wary of those who will tell you that you have to change the answers in the model to fit existing biases of senior management or to "play politics." For organizations to accept the results of the process, those results must be integral and acknowledged as having integrity, not just being the news that senior management wants to hear.

That said, not all of your stakeholders will need to understand the ins and outs of the model. A peer review of the model for technical correctness is strongly recommended, but beyond that you need to focus on answering the actual questions that stakeholders have, not just telling them all about how your model is the best thing since sliced bread. It is also important to communicate the sensitivity of the model to key assumptions and conditions.

OBJECTIVE 2. DELIVER REPORT WITH THE FINDINGS, or OBJECTIVE 3. CREATE MODEL, USABILITY, & SYSTEM REQUIREMENTS FOR PRODUCTION

Your report format will vary with the organization and how it will use your report. The main thing is that the report needs to have a clear message. Either recommend a course of action or no action and state your reasons. Basic report guidelines apply to this as with any other report. An executive summary and recommendations for further action should be at the front, with the supporting details, methodology, and references for further knowledge in the main body. Be clear about the assumptions and limitations of your model. Your audience should have enough information to judge whether the model you have developed meets the needs of the project, or where future resources should be directed. Use graphical aids to communicate

findings whenever possible. Well-constructed graphics can simplify results and uncover patterns that are easily missed in tables. Always consider good graphical practice. A poorly constructed graphic can be misleading, as outlined very well in books by Tufte (2001), Few (2012), among others.

OBJECTIVE 5. SUPPORT DEVELOPMENT

Two key items to consider as a model becomes the basis for an organization taking action are:

- Plan the deployment – This task develops a strategy for deployment.

- Plan monitoring and maintenance – A detailed monitoring plan is needed to ensure the results are being used correctly and to determine any problems with the model or data.

Today, most analytics are deployed within an organization's business processes. They will receive inputs from the business process, analyze the data in real time, and produce a result that is fed back into the business process. For example, in a customer relationship management (CRM) system for a cell phone carrier, a customer call to the call center may trigger an analysis of the customer's likelihood to cancel their service ("churn analysis") based on their demographics and bill history. This information may trigger an alert to the call center operator to provide a special offer to the customer.

Deploying within a business process requires an understanding of both the analytics and the business process. You must identify where in the process the analytics will be triggered and what data will need to be passed to the analytics and back to the business process. Then determine the actions in the business process that should be taken as a result of the analytics.

Periodically survey and interview key stakeholders to see how their day-to-day interaction with the model is going and how their results have changed since they started using the model. Pay particular attention to functional areas where the model is being ignored as irrelevant, as they'll tell you where key assumptions either have been already invalidated or will soon become invalidated and use that as a way to strengthen and update the model.

The documentation can be minimal to extensive depending on how complex the deployment will be. Often forgotten is the training documentation. This will range from a minimal effort, if the solution is designed for your fellow analysts, all the way to extensive in-depth training if fundamental business processes are changing as a result of the new business model.

SUMMARY

Deployment is a critical step to make the analytics actionable. Proper preparation will allow for a smooth deployment that ensures the success of the analytics efforts. Further information is available from Chapman et al. in the CRISP-DM methodology.

FURTHER READING

Chapman P, et al., CRISP-DM 1.0 Step by Step data mining guide, http://lyle. smu.edu/~mhd/8331f03/crisp.pdf and http://www.the-modeling-agency.com/crisp-dm.pdf.

Few S (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, 2nd ed. (Analytics Press, Burlingame, CA).

Laursen GHN, Thorlund J (2010) *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting* (John Wiley & Sons, Hoboken, NJ).

Tufte ER (2001) *The Visual Display of Quantitative Information*, 2nd ed. (Graphics Press, Cheshire, CT).

Analytical models will go through a lifecycle from conception and business justification through the model building and deployment. A good lifecycle process helps to keep this process orderly, minimizes the cost and efforts of creating and maintaining the models, and, importantly, provides the business users with clear roles within the lifecycle.

Now is the time to think through the process you want to define for building and deploying analytics, before the deadlines of the business require this to be done in an ad hoc manner. An effective process requires defining the roles of the various departments involved and the governance process that will be used to iron out differences and make decisions.

Learning Objectives

1. Document initial structure

2. Track model quality

3. Recalibrate and maintain the model*

4. Support training activities

5. Evaluate the business benefit of the model over time

OBJECTIVE 1. DOCUMENT INITIAL STRUCTURE

It can be tempting during the rush of data collection and model building to skimp on documentation, figuring that there will be time to write it down later once things settle down. Do not fall for this. People will inevitably leave the project before completing their documentation if you do. For the model to be trusted it has to be repeatable, and that means writing down what you and your team did and how you did it.

Documentation should include at least the following:

• Key assumptions made about the business context and analytics problem

• Data sources and data schema

• Methods used to clean and harmonize the data

• Model approach and model review artifacts

*Tasks performed by analytics professionals beyond CAP® certification level

- Documentation for any software code written

- Recommendations for future improvements to the model

Essentially you are leaving behind enough of a record for someone else to come in and recreate the model and get the same results. This documentation should be kept in a known place, ideally backed up in a few different places.

OBJECTIVE 2. TRACK MODEL QUALITY

Evaluation criteria should be created up front both in terms of the business results expected and the accuracy and confidence expected from the model. Some of the criteria that might be used include:

- Value of the model in terms of the business.

- Does the model discover/predict something that is new and useful?

- Is the model reliable across a wide range of data?

- Can a "lift" or "gain" graph be constructed to show how well the model is predicting?

- Check if the model's predictions on unknown data are as good as the predictions on the data that were used to train or build the model.

The model should be routinely checked over time and quality parameters recorded. When the model quality starts to decay, it is time for the next step of recalibrating the model and rechecking its assumptions.

OBJECTIVE 3. RECALIBRATE & MAINTAIN THE MODEL

The results from the model should be tracked over the long term because even a model that performs well initially may degrade as input data changes or user requirements change. Additionally, the model results may also help in areas beyond that expected, such as identifying data quality problems, or new areas for modeling. In the case of data quality problems or minor changes in the business environment, a simple recalibration of the model, similar to what was done in the model building phase of partitioning data into training and testing data, and constructing the parameters will be sufficient to get the model working again. If there has been a fundamental change in a key assumption or two, however, then the project needs to be revalidated against the business problem to see if the overall approach is even still valid. No model lasts forever. At some point the resulting model will need to be improved, replaced, or sunset.

OBJECTIVE 4. SUPPORT TRAINING ACTIVITIES

One of the keys to a successful analytics project or engagement is appropriate training for the users of the model and its results. While the training does not need to cover all the intricacies of the model, the training should ensure that the users understand the business use of the analytics model and how to interpret the results. As the analyst, you must also ensure that the users do not conclude more from the results than the model is capable of producing.

OBJECTIVE 5. EVALUATE THE BUSINESS BENEFIT OF THE MODEL OVER TIME

As your analytics effort takes shape and grows within your organization, you will be fighting for resources to do more and better projects. A key weapon in that fight is being able to point to the benefits that your previous models have brought to the organization. How much money has the organization made because your models pointed the way? How much money has the organization saved because your models pointed out wasted effort? To answer these questions in a defensible manner, you have to be able to evaluate the business benefit of the model over time. To do that, you need to be able to simulate what the organization would have been doing without the changes wrought by the model.

One way to do that is by looking at how your organization is doing against industry benchmarks during the time period in question. Have you grown from a second quintile organization to a first quintile in a key area? Another way is to look at how products that have been modeled have changed their financial returns to the organization. Has net profit grown since the model was introduced? How about return on net assets?

Whatever way you approach it, evaluating the business benefit allows you to "keep score" and market your capabilities to the organization at large, helping it grow and develop by solving business problems that are otherwise insoluble.

SUMMARY

This chapter presented an overview of the lifecycle for an analytics project. Whether the practitioner follows this exact methodology or develops their own, it is important for project success to have a defined methodology that can be used from project to project. This allows a team of analytics professionals that perhaps have not worked together before to quickly come together, easily communicate, and deliver professional results in a timely manner. A defined process based on best practices and lessons learned will also help avoid common problems such as skipping an important step.

Chapman P, et al., CRISP-DM 1.0 Step by Step data mining guide, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf and http://www.the-modeling-agency.com/crisp-dm.pdf.

Wirth R (2000) CRISP-DM: Towards a standard process model for data mining. *Proc. Fourth Internat. Conf. Practical Appl. Knowledge Discovery Data Mining,* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133.

Introduction

Not only does a professional analyst need mathematical ability, statistical knowledge, data skills, etc., he or she also needs to possess those softer skills related to communication and understanding. Without the ability to convince or explain the problem, problem solution, and implications, an entire analytics project can be jeopardized.

Communication skills are as necessary to the analytics professional as math and reasoning skills. Although many analytics are among the less extroverted, that does not mean they are less capable or are less focused on ensuring success. Here you will learn why the softer skills are an important component of the analytics professional profile. An analytics professional needs the following:

- Ability to communicate with a client/employer regarding the framing of an analytics problem.

- Understanding the background of the client/employer regarding its organization and specific industry focus.

- Ability to explain the findings of the analytics process in sufficient detail to ensure clear understanding by the client/employer.

Learning Objectives

By the end of this chapter you should be able to:

1. Recognize the importance of soft skills

2. Determine the need to communicate with stakeholders

3. Determine how to tailor communication so as to be understood

TASK 1: TALKING INTELLIGIBLY WITH STAKEHOLDERS WHO ARE NOT FLUENT IN ANALYTICS

For consistency of organization, we are listing soft skills as tasks. However, these are not tasks to be completed on a deadline; the need for these skills is consistent throughout the entire project. Domain I of the CAP® program focuses on framing the business problem, and Domain II on determining whether the business

problem has an analytics solution. Here we focus on communicating the content of Domains I and II to external stakeholders, those who may be less versed in the process.

For example, if you are approached by a client or employer who states that sales of their industry-leading widget are falling and they want to know how to optimize the pricing structure, your first response should not be "yes, of course" or "no, this can't be done." Your first response is to engage the client in a dialog to discover what they really want. Do they want to find out why sales are falling? Do they have reason to believe changing the pricing is the best response? What data do they have on past sales, customers, supply chain, commodities pricing—all of which could have an impact on sales figures. If price is the only thing they are concerned with, it doesn't matter what they're selling or to whom. As Seth Godin (2013) wrote in his blog post "Q&A: Purple Cows and Commodities" (http://sethgodin.typepad.com/seths_blog/2013/08/qa-purple-cows-and-commodities.html), "If you tell me that price is the only thing that matters to customers, I respond that nothing about this product matters to them."

The job of an analytics professional is to find the deep underlying motives of any client engagement: he or she is (almost) an analyst in the mode of a psychiatric worker. Question, question, question until it is clear what the problem is and how a solution can be attempted.

The mechanisms of communication are well studied; we like to communicate with those who share our world view. We dislike communicating with those who don't and often with those who try to reshape our views. The analytics professional is sometimes in the position of reshaping viewpoints. As previously noted, when the sales of an item decrease and the client says we want to optimize the price, they may not be amenable to hearing that regardless of price, the item is not valued, or that their marketing has targeted a customer base unlikely to purchase, or any of the many other reasons analytics might uncover.

For successful interactions in such cases, it may be helpful for the analytics professional to unleash their inner four-year-old and keep asking 'Why?' This should be done with care—it doesn't take long to get impatient at a long string of why's. Rather, the professional learns to reframe the answer to a question in such a way that it continues to drill down to the answer. Learning this skill can save time, money, and project success.

Victoria A. Williams (2011) notes in her dissertation titled "Effective communication techniques for eliciting information technology requirements" that IT project failure costs businesses millions of dollars. She attributes those failures and the associated cost to the lack of communication between members of the business and technology communities. One of the major problems as identified by many

researchers is the limitations of human factors otherwise known as the softer skills. Complexities arise because each party in the communication exchange views the exchange through different conceptual frameworks; this difference in perspectives wedges a gap between what is being communicated by one party and heard by the other.

Communications techniques include asking open ended questions; instead of asking "what did you do last week when the system crashed?" ask "what do you think is the best option to counteract the losses incurred when a system crashes?" That question takes things out the realm of what you did last week into the more theoretical realm of what is the best thing to do. From there, it is much easier to direct a conversation toward best practices.

Empathy is not in and of itself a communication technique, but it certainly does enhance communication. Being empathetic doesn't mean you have to like someone, it just means you have to establish a human link. At some point, we are all linked and we all have families and/or teammates and/or pets and/or hobbies and/or books we're reading and/or musicians we listen to. Just noticing something about the person you're addressing in a neutral, non-value-laden way is a good opener to a conversation. For instance, don't lead off with anything related to politics or religion; it is much safer to comment on the weather, which is a common experience whether it happens to be raining or the sun is shining. The ultimate goal is the success of the analytics project, not the unfriendliness of the project sponsor or the end users. Common courtesy can often be used here: shake hands, introduce team members, offer coffee or tea, say please and thank you. As Daniel Pink notes in his recent book To Sell is Human, we're all salesmen to some degree.

TASK 2: CLIENT/EMPLOYER BACKGROUND & FOCUS

Concurrently with determining the problem, a professional analyst determines the client or employer's background and focus. Where does the client (for purposes of brevity we will refer to the client/employer as the client) work within the organization? Knowing that gives the analyst information. For example, if the client is in the IT department, he or she will likely want a solution that uses the IT infrastructure to its full capacity. If the client is in the profit and loss department, he or she will want a solution that maximizes profit and minimizes loss. If the client is in the marketing department, he or she will want a solution that delights the customer. Sometimes the solution cannot be optimized for everyone; an elegant solution will take into account the concerns of all client stakeholders. Again, the analytics professional needs to enhance his or her use of communication. Asking for an organizational chart helps but may not suffice; inter- and intraoffice communications often follow an informal chain that is outside the organization chart. Take note of the people in

the project management meetings; their presence may be an indication of their status within the organization.

Determine who the project stakeholders are—it is likely more than one person. It could be the C-level who want to optimize the bottom line; it might be the IT people who want to optimize use of their systems; it might be the owner of process; it might be the workers who will use the new process: in short, it could be anyone from the highest to the lowest levels of the organization and perhaps even include the customers who are outside the organization, and often is all of the above.

Eliciting the needs of all stakeholders is essential. Once elicited, however, the need must be prioritized to enhance the solution. If the solution requires the purchase of new software or the collection of new data, there are myriad ways to present profit and loss along a sliding scale, or if the customer is not delighted by all the background work, then communication is really essential to explain why these things happened or will happen in such a way as to exclude no one.

TASK 3: CLARIFYING THE ANALYTICS PROCESS

The analytics professional is the person at the heart of the analytics process. That person has an understanding of the entire process from beginning to life-cycle maintenance. He or she may be engaged to work with a client who is not versed in the process but has heard that analytics is a wonderful tool to promote one's business. It is the job of the analytics professional to ensure that his or her questions and comments are seen as necessary to the process, not as intrusive and time wasting.

Many are familiar with ISO 9000 series accreditation. The process of obtaining that accreditation is massive and intrusive, time consuming and painstaking, etc. However, if all concerned are aware that the end is to ensure that they contribute to this demonstration of excellence, it becomes less of an intrusion and is seen more as a contribution to an overall process. If this is not made clear, then individuals may be less willing to spend time enumerating the processes they follow and will slow the entire application process. This could in turn mean a less viable response to requests for proposals, which could mean fewer dollars of income at the corporate level and the loss of a job. There are those who may not see the dire progression of events unless it is laid out clearly, which in the case of the analytics process, the professional is sure to do.

Not only should the entire process be transparent to all involved, but there are times when the analytics professional is called on to be a translator. He or she must be able to move from very technical fields with associated jargon and acronyms to

a less technical field where there is little or no familiarity with specific terminology related to the analytics process. Were the average client or stakeholder familiar with the terminology and the steps to the process, the analytics professional might find him or herself to be an extraneous expense rather than an added value after all.

SUMMARY

Not only does the analytics professional need to have knowledge of the analytics process, he or she must have sufficient command of the less science-based skill set that enables easy communication and coordination with stakeholders, clients, and users. The analytics professional should be agile and able to move easily from the technical to the nontechnical areas of an analytics project, and relay to each sector the perspective of the other.

FURTHER READING

Godin S (2013) Q&A: Purple Cows and Commodities, http://sethgodin.typepad. com/seths_blog/2013/08/qa-purple-cows-and-commodities.html.

The Ladder of Inference: Avoiding "Jumping to Conclusions", http://www. mindtools.com/pages/article/newTMC_91.htm.

Pink D (2013) *To Sell is Human: The Surprising Truth about Moving Others* (Riverhead books, New York).

Timmer J (2013) Applying science to communicate science: Right now, it's hard to find relevant information on how to do it well, August 1, http://arstechnica.com/ staff/2013/08/applying-science-to-communicate-science/.

Weinschenk SM (2013) *How to Get People to Do Stuff: Master the Art and Science of Persuasion and Motivation* (Peachpit, San Francisco).

Williams VA (2012) Effective communication techniques for eliciting information technology requirements, ProQuest, UMI Dissertation Publishing, Ann Arbor, MI.

THIS PAGE IS DELIBERATELY LEFT BLANK

Every individual prepares for an exam in their own way. That having been said and recognizing that education and experience differ for most, we suggest that the following may be of help.

First, cultivate good study habits: most of us have done this over our academic career, but sometimes we forget that studying is not quite the same as reading and tend to sit in our easy chairs and half-heartedly listen to TV and study. That might work for some but is generally not the best way to do it. Study in a quiet place with minimal distraction; take notes, whether those are penciled in the margin or highlighted on a computer screen. Ask yourself questions: on a scale of (1) 'Never heard of this before' to (10) 'I could teach this in my sleep', where does your knowledge lie? If it is closer to 1, then maybe you need to find out more; if it's closer to 10, then maybe you don't need more study and can help others.

We suggest that studying in a group is more conducive to learning than studying alone because in a group, you get the benefit of not only your own experience, but that of others as well. The thing to watch out for in a group is too much time spent sharing best practices (aka war stories) as opposed to exploring the subject matter.

A study group can be organized formally or informally; meeting space can be in a conference room or at a local coffee shop or online. If you set aside 6 weeks to meet once a week for an hour, you can divide the labor and have one person read ahead and/or research a topic area and share with the others (best practices should be limited to those practices dealing directly with the topic area). For an example of a six week study group schedule, see below. This can be adjusted depending on whether you're studying in a group of six or going it solo.

No matter whether you study alone or in a group, do remember to check the review questions and rather than selecting the correct answer, challenge yourself to explain why it is correct and why the other answers are not correct. Another good practice is to talk to people who have already taken the test and get their insights on preparation. Finally, remember that the exam is based on what you as an analytics professional do, so use your experience, common sense and adhere to the highest ethical practices of analytics professionals. Use this Study guide as a general guide but not necessary the only material to ensure your success on the exam.

Week 1: Business problem framing and analytic problem framing

- What's the business problem: too little revenue, too slow process, too many returns, too few customers? Are these problems that have an analytics solution or is there a different solution?

Week 2: Data concerns

- What data are available? What data are needed? Are the data usable in their present format? What needs to change in the data collection or data format to implement an analytics solution?

Week 3: Methodologies for problem solution

- What methodology will be most useful to solve the problem? Irrespective of the tool used, what methodology will provide the optimum solution?

Week 4: Models using those methodologies

- What models can use the identified methodology to aid in solution of the analytics problem? Can you build a model to take advantage of the methodology?

Week 5: Deploying and using the model

- What do you need to deploy the model? Who will be using the model? Who will need to get the results?

Week 6: Lifecycle maintenance & wrap up

- How will you know that the model is still providing the same solution to the original problem? How can you tell when or if the data are no longer providing the agreed upon results? What do you do if the model provides skewed data?

GLOSSARY

TERM	DEFINITION
5S	workplace organization method promoting efficiency and effectiveness; five terms based on Japanese words for: sorting, set in order, systematic cleaning, standardizing, and sustaining (http://en.wikipedia.org/wiki/5S_(methodology)
5 Whys	iterative process of discovery through repetitively asking 'why'; used to explore cause and effect relationships underlying and/or leading to problem (http://en.wikipedia.org/wiki/5_Whys)
80/20 Rule	AKA the Pareto principle: roughly 80% of results come from 20% of effort
Accuracy	quality or state of being correct or precise, or the degree to which the result of a measurement, calculation, or specification conforms to the correct value or standard (https://www.google.com/#q=accuracy)
Activity-based costing	method of assigning costs to products or services on the resources that they consume (http://www.economist.com/node/13933812)
Agent-based modeling	a class of computation models for simulating actions and interactions of autonomous agents with a view to assessing their effects on the system as a whole (http://en.wikipedia.org/wiki/Agent-based_model)
Algorithm	set of specific steps to solve a problem

Amortization	allocation of cost of an item or items over a time period such that the actual cost is recovered; often used to account for capital expenditures
Analytics	scientific process of transforming data into insight for making better decisions (INFORMS)
Analytics professional	person capable of making actionable decisions through the analytic process; also a person holding the Certified Analytics Professional (CAP®) credential
ANCOVA	acronym for analysis of covariance
ANOVA	acronym for analysis of variance
Artificial Intelligence (AI)	branch of computer science that studies and develops intelligent machines and software (http://en.wikipedia.org/wiki/Artificial_intelligence)
Artificial Neural Networks	computer-based models inspired by animal central nervous systems (https://www.google.com/#q=artificial+neural+networks)
Assemble-to-Order (ATO)	manufacturing process where products are assembled as they are ordered; characterized by rapid production and customization (http://www.investopedia.com/terms/a/assemble-to-order.asp)
Assignment problem	one of the fundamental combinatorial optimization problems in the branch of optimization or operations research in mathematics; consists of finding a maximum-weight matching in a weighted bipartite graph (http://en.wikipedia.org/wiki/Assignment_problem)

Automation	use of mechanical means to perform work previously done by human effort
Average	sum of a range of values divided by the number of values to arrive at a value characteristic of the midpoint of the range; see also, Mean
Batch production	method of production where components are produced in groups rather than a continual stream of production; see also, Continuous production
Benchmarking	act of comparison against a standard or the behavior of another in attempt to determine degree of conformity to standard or behavior
Benchmark problems	comparison of different algorithms using a large test set (http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume24/ortizboyer05a-html/node6.html)
Bias	a tendency for or against a thing, person, or group in a way as to appear unfair; in statistics, data calculated so that it is systematically different from the population parameter of interest (http://en.wikipedia.org/wiki/Bias_(statistics)
Big data	data sets too voluminous or too unstructured to be analyzed by traditional means
Box-and-whisker plot	a simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle (http://oxforddictionaries.com/us/definition/english/box-plot)

Branch-and-Bound	a general algorithm for finding optimal solutions of various optimization problems; consists of a system enumeration of all candidate solutions where large subsets of fruitless candidates are discarded en masse using upper and lower estimated bounds of the quantity being optimized (http://en.wikipedia.org/wiki/Branch_and_bound)
Business analytics (BA)	refers to the skills, technologies, applications, and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning; can be descriptive, prescriptive, or predictive; focuses on developing new insights and understanding of business performance based on data and statistical methods (http://en.wikipedia.org/wiki/Business_analytics and www.informs.org)
Business case	reasoning underlying and supporting the estimates of business consequences of an action
Business intelligence (BI)	a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information (http://en.wikipedia.org/wiki/Business_intelligence)
Business Process Modeling or Mapping (BPM)	act of representing processes of an enterprise so that the current process may be analyzed and improved; typically action performed by business analysis and managers seeking improved efficiency and quality (http://en.wikipedia.org/wiki/Business_process_modeling)

Chief Analytics Officer (CAO)	possible title of one overseeing analytics for a company; may include mobilizing data, people, and systems for successful deployment, working with others to inject analytics into company strategy and decisions, supervising activities of analytical people, consulting with internal business functions and units so they may take advantage of analytics, contracting with external providers of analytics (Davenport, Enterprise Analytics, p. 173)
Chi-squared Automated Interaction Detection (CHAID)	a technique for performing decision tree analysis developed by Gordon V. Kass. CHAID is one of several commonly used techniques for decision trees and is based upon hypothesis testing using Bonferroni correction.
Classification	assortment of items or entities into predetermined categories
Cleansing	AKA cleaning or scrubbing: the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database; may also involve harmonization of data, and standardization of data (http://en.wikipedia.org/wiki/Data_cleansing)
Clustering	grouping of a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups or clusters (http://en.wikipedia.org/wiki/Cluster_analysis)
Combinatorial optimization	a topic that consists of finding an optimal object from a finite series of objects; used in applied mathematics and theoretical computer science (http://en.wikipedia.org/wiki/Combinatorial_optimization)

Confidence interval	a type of interval estimate of a population parameter used to indicate the reliability of an estimate. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest if the experiment is repeated (http://en.wikipedia.org/wiki/Confidence_interval)
Confidence level	if confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will match the confidence level (http://www.usablestats.com/lessons/ConfidenceLevel)
Conjoint analysis	allows calculation of relative importance of varying features and attributes to customers
Constraint	a condition that a solution to an optimization problem is required by the problem itself to satisfy. There are several types of constraints—primarily equality constraints, inequality constraints, and integer constraints (http://en.wikipedia.org/wiki/Constraint_(mathematics))
Constraint programming	a programming paradigm wherein relations between variables are stated in the form of constraints (http://en.wikipedia.org/wiki/Constraint_programming)
Continuous production	method of production where components are produced in a continuous stream; see also, Batch production
Correlation	a broad class of statistical relationships involving dependence (http://en.wikipedia.org/wiki/Correlation_and_dependence)

Cost of capital	the cost of funds used for financing a business. Cost of capital depends on the mode of financing used—it refers to the cost of equity if the business is financed solely through equity, or to the cost of debt if it is financed solely through debt (www.investopedia.com)
Cube	see OLAP cube
Cumulative density function	probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x; used to specify the distribution of multivariate random variables (http://en.wikipedia.org/wiki/Cumulative_distribution_function)
Cutting stock problem	optimization or integer linear programming problem arising from applications in industry where high production problems exist (http://en.wikipedia.org/wiki/Cutting_stock_problem)
Data	(plural form of datum) values of qualitative or quantitative variables, belonging to a set of items; represented in a structure, often tabular (represented by rows and columns), a tree (a set of nodes with parent-children relationship), or a graph structure (a set of interconnected nodes); typically the results of measurements (http://en.wikipedia.org/wiki/Data)
Data mining	relatively young and interdisciplinary field of computer science; the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems; see also, KDD (Davenport, Enterprise Analytics, p. 14)

Data warehouse	a central repository of data that is created by integrating data from one or more disparate sources; used for reporting and data analysis (http://en.wikipedia.org/wiki/Data_warehouse)
Database	an organized collection of data organized to model relevant aspects of reality to support processes requiring this information (http://en.wikipedia.org/wiki/Database)
Decision tree	graphic illustration of how data leads to decision when branches of the tree are followed to their conclusion; different branches may lead to different decisions
Decision variables	a decision variable represents a problem entity for which a choice must be made. For instance, a decision variable might represent the position of a queen on a chessboard, for which there are 100 different possibilities (choices) on a 10x10 chessboard or the start time of an activity in a scheduling problem. Each possible choice is represented by a value, hence the set of possible choices constitutes the domain that is associated with a variable (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Descriptive analytics	prepares and analyzes historical data to identify patterns for reporting trends (http://www.informs.org/Community/Analytics/About-Us)
Design of experiments	design of any information gathering exercise where variation is present, whether under the control of the experimenter or not; see also, Experimental design (http://en.wikipedia.org/wiki/Design_of_experiments)

Discrete event simulation	models the operation of a system as a discrete sequence of events in time; between events, no change in the system is assumed thus a simulation can move in time from one event to the next (http://en.wikipedia.org/wiki/Discrete_event_simulation)
Dynamic programming	based on the Principle of Optimality, this was originally concerned with optimal decisions over time. For continuous time, it addresses problems in variational calculus. For discrete time, each period is sometimes called a stage, and the DP is called a multistage decision process. Here is the Fundamental Recurrence Equation for an additive process: $F(t, s) = Opt\{r(t, s, x) + aF(t', s')$: x in X(t, s) and s'=T(t, s, x)\}, (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Effective domain	the domain of a function for which its value is finite (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Efficiency	the comparison of what is actually produced or performed with what can be achieved with the same consumption of resources (money, time, labor, etc.). It is an important factor in determination of productivity (www.businessdictionary.com)
Engagement	an estimate of the depth of visitor interaction against a clearly defined set of goals; may be measured through analytical models (Davenport, Enterprise Analytics, p. 73-74)

Enterprise resource planning (ERP)

a cross-functional enterprise system driven by an integrated suite of software modules that supports the basic internal business processes of a company (http://en.wikipedia.org/wiki/Enterprise_resource_planning)

ETL (extract, transform, load)

refers to three separate functions combined into a single programming tool. First, the extract function reads data from a specified source database and extracts a desired subset of data. Next, the transform function works with the acquired data—using rules or lookup tables, or creating combinations with other data—to convert it to the desired state. Finally, the load function is used to write the resulting data (either all of the subset or just the changes) to a target database, which may or may not previously exist (http://searchdatamanagement.techtarget.com/definition/extract-transform-load)

Experimental design

in quality management, a written plan that describes the specifics for conducting an experiment, such as which conditions, factors, responses, tools, and treatments are to be included or used; see also, Design of experiments (www.businessdictionary.com)

Expert systems

a computer program that simulates the judgment and behavior of a human or an organization that has expert knowledge and experience in a particular field. Typically, such a system contains a knowledge base containing accumulated experience and a set of rules for applying the knowledge base to each particular situation that is described to the program (http://searchcio-midmarket.techtarget.com/definition/expert-system)

Factor analysis

a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Factor analysis searches for such joint variations in response to unobserved latent variables (http://en.wikipedia.org/wiki/Factor_analysis)

Failure Mode and Effects Analysis (FMEA)	a systematic, proactive method for evaluating a process to identify where and how it might fail, and to assess the relative impact of different failures to identify the parts of the process that are most in need of change (http://intranet.uchicago.edu/quality/FailureModesandEffectsAnalysis_FMEA_1.pdf)
Fixed cost	a cost that is some value, say C, regardless of the level as long as the level is positive; otherwise the fixed charge is zero. This is represented by Cv, where v is a binary variable. When v = 0, the fixed charge is 0; when v = 1, the fixed charge is C. An example is whether to open a plant (v = 1) or not (v = 0). To apply this fixed charge to the non-negative variable x, the constraint x <= Mv is added to the mathematical program, where M is a very large value, known to exceed any feasible value of x. Then, if v = 0 (e.g., not opening the plant that is needed for x > 0), x = 0 is forced by the upper bound constraint. If v = 1 (e.g., plant is open), x <= Mv is a redundant upper bound. Fixed charge problems are mathematical programs with fixed charges (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Forecasting	the use of historic data to determine the direction of future trends (http://www.investopedia.com/terms/f/forecasting.asp)
Fuzzy logic	a form of mathematical logic in which truth can assume a continuum of values between 0 and 1 (http://wordnetweb.princeton.edu/perl/webwn?s=fuzzy logic)

Game Theory   in general, a (mathematical) game can be played by
                       one player, such as a puzzle, but its main connection
                       with mathematical programming is when there are at
                       least two players, and they are in conflict. Each player
                       chooses a strategy that maximizes his payoff. When
                       there are exactly two players and one player's loss is the
                       other's gain, the game is called zero sum. In this case, a
                       payoff matrix A is given where A_{ij} is the payoff to player
                       1, and the loss to player 2, when player 1 uses strategy
                       i and player 2 uses strategy j. In this representation
                       each row of A corresponds to a strategy of player 1,
                       and each column corresponds to a strategy of player 2.
                       If A is m × n, this means player 1 has m strategies, and
                       player 2 has n strategies (A. Holder, editor. Mathematical
                       Programming Glossary. INFORMS Computing Society,
                       http://glossary.computing.society.informs.org/, 2006-08.
                       Originally authored by Harvey J. Greenberg, 1999-2006.)

Genetic algorithms   a class of algorithms inspired by the mechanisms of
                       genetics, which has been applied to global optimization
                       (especially for combinatorial programs). It requires
                       the specification of three operations (each is typically
                       probabilistic) on objects, called "strings" (A. Holder,
                       editor. Mathematical Programming Glossary. INFORMS
                       Computing Society, http://glossary.computing.society.
                       informs.org/, 2006-08. Originally authored by Harvey J.
                       Greenberg, 1999-2006.) Originally authored by Harvey J.
                       Greenberg, 1999-2006.)

Global optimal	refers to mathematical programming without convexity assumptions, which are NP-hard. In general, there could be a local optimum that is not a global optimum. Some authors use this term to imply the stronger condition there are multiple local optima. Some solution strategies are given as heuristic search methods (including those that guarantee global convergence, such as branch and bound). As a process associated with algorithm design, some regard this simply as attempts to assure convergence to a global optimum (unlike a purely local optimization procedure, like steepest ascent). (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006. See the supplement by J.D. Pintér.)
Goodness of fit	degree of assurance or confidence to which the results of a sample survey or test can be relied upon for making dependable projections. Described as the degree of linear correlation of variables, it is computed with the statistical methods such as chi-square test or coefficient of determination (www.businessdictionary.com)
Graphical User Interface (GUI)	a human–computer interface (i.e., a way for humans to interact with computers) that uses windows, icons, and menus, and that can be manipulated by a mouse (and often to a limited extent by a keyboard as well) (http://www.linfo.org/gui.html)
Greedy heuristics	an algorithm that follows the problem-solving heuristic of making the locally-optimal choice at each stage with the hope of finding a global optimum (http://en.wikipedia.org/wiki/Greedy_heuristic)

Heuristic	in mathematical programming, this usually means a procedure that seeks an optimal solution but does not guarantee it will find one, even if one exists. It is often used in contrast to an algorithm, so branch and bound would not be considered a heuristic in this sense. In AI, however, a heuristic is an algorithm (with some guarantees) that uses a heuristic function to estimate the "cost" of branching from a given node to a leaf of the search tree (Also, in AI, the usual rules of node selection in branch and bound can be determined by the choice of heuristic function: best-first, breadth-first, or depth-first search) (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Histogram	graphic depiction of data using columns to represent relative size/importance of data grouping
Hypothesis testing	the theory, methods, and practice of testing a hypothesis by comparing it with the null hypothesis. The null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance (https://www.google.com/#psj=1&q=hypothesis+testing+definition)
Influence diagram	depicts structure of decision process and notes the data needed to make the decision
INFORMS	the largest professional society in the world for professionals in the field of operations research (OR), management science, and analytics (www.informs.org/About)

Innovative Applications in Analytics Award	award administered by the Analytics Section of INFORMS to recognize creative and unique developments, applications, or combinations of analytical techniques. The prize promotes the awareness of the value of analytics techniques in unusual applications, or in creative combination to provide unique insights and/or business value (http://www.informs.org/Community/Analytics/News-Events2/Innovative-Applications-in-Analytics-Award)
Integer program	the variables are required to be integer-valued. Historically, this term implied the mathematical program was otherwise linear, so one often qualifies a nonlinear integer program versus a linear IP (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Integrity	the measure of the trust that can be placed in the correctness of the information supplied by a navigation system (http://www.navipedia.net/index.php/Integrity; http://www.genengnews.com/gen-articles/preserving-the-integrity-of-statistics/3081/)
Internal rate of return (IRR)	the rate of growth that a project or investment is expected to create, expressed as a percentage, over a specified term. IRR is, in essence, the theoretical interest rate earned by the project (http://www.askjim.biz/answers/internal-rate-of-return-irr-definition_4754.php)
KDD	acronym for knowledge discovery in databases process; see also, Data mining (Davenport, Enterprise Analytics, p. 14)

Knapsack problem | an integer program of the form, Max{cx: x in Z^{n+} and ax <= b}, where a > 0. The original problem models the maximum value of a knapsack that is limited by volume or weight (b), where x_j = number of items of type j put into the knapsack at unit return c_j, that uses a_j units per item (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary. computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)

Lead time | time between the initial phase of a process and the emergence of results, as between the planning and completed manufacture of a product (http://www. thefreedictionary.com/lead+time)

Lean production | a Japanese approach to management that focuses on cutting out waste while ensuring quality. This approach can be applied to all aspects of a business – from design through production to distribution (http://www. tutor2u.net/business/production/introduction-to-lean-production.html)

Lift/lift curve | a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model; lift charts consisting of lift curve and a baseline are visuals aids for measuring model performance (http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html)

Linear program	opt{cx: Ax = b, x >= 0}. (Other forms of the constraints are possible, such as Ax <= b.) The standard form assumes A has full row rank. Computer systems ensure this by having a logical variable (y) augmented, so the form appears as Opt{cx: Ax + y = b, L <= (x, y) <= U} (also allowing general bounds on the variables). The original variables (x) are called structural. Note that each logical variable can be a slack, surplus, or artificial variable, depending on the form of the original constraint. This computer form also represents a range constraint with simple bounds on the logical variable. Some bounds can be infinite (i.e., absent), and a free variable (logical or structural) is when both of its bounds are infinite (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Little's Law	queuing theory where numerator and denominator are halved so queues are roughly equivalent no matter how many are in line; the long-term average number of customers in a stable system L is equal to the long-term average effective arrival rate, λ, multiplied by the (Palm) average time a customer spends in the system, W; or expressed algebraically: $L = \lambda W$. The relationship is not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else. (http://en.wikipedia.org/wiki/Little's_law)
Local optimal	a solution that is optimal (either maximal or minimal) within a neighboring set of candidate solutions (http://en.wikipedia.org/wiki/Local_optimum)

Logistic regression

a type of probabilistic classification model [1] used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features). Logistic regression can be binomial or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (for example, "dead" versus "alive"). Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "better" versus "no change" versus "worse") (http://en.wikipedia.org/wiki/Logistic_regression)

Machine learning

an artificial intelligence (AI) discipline geared toward the technological development of human knowledge. Machine learning allows computers to handle new situations via analysis, self-training, observation, and experience (http://www.techopedia.com/definition/8181/machine-learning)

MANOVA

acronym for multivariate analysis of variance for use with multiple independent variables

Mean

the arithmetic average of a set of values or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode); see also, Average (http://en.wikipedia.org/wiki/Mean)

Mean squared error (MSE)

the unbiased estimator of population variance. MSE divides by the error degrees of freedom, e.g., if only the mean is estimated, MSE divides by N-1, if four parameters are estimated, MSE divides by N-4, and so on (http://en.wikipedia.org/wiki/Mean_squared_error)

Mean time between failures (MTBF)	a measure of how reliable a hardware product or component is. For most components, the measure is typically in thousands or even tens of thousands of hours between failures (http://whatis.techtarget.com/definition/MTBF-mean-time-between-failures)
Median	the value such that the number of terms having values greater than or equal to it is the same as the number of terms having values less than or equal to it (http://searchdatacenter.techtarget.com/definition/statistical-mean-median-mode-and-range)
Metaheuristics	a general framework for heuristics in solving hard problems. The idea of ``meta'' is that of level. An analogy is the use of a metalanguage to explain a language. For computer languages, we use symbols, like brackets, in the metalanguage to denote properties of the language being described, such as parameters that are optional. Examples of metaheuristics are: Ant Colony Optimization, Genetic Algorithms, Memetic Algorithms, Neural networks, etc. (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Mode	value of the term that occurs the most often (http://searchdatacenter.techtarget.com/definition/statistical-mean-median-mode-and-range)
Monte Carlo simulation	a computerized mathematical technique that allows people to account for risk in quantitative analysis and decision making. The technique is used by professionals in such widely disparate fields as finance, project management, energy, manufacturing, engineering, research and development, insurance, oil and gas, transportation, and the environment (http://www.palisade.com/risk/monte_carlo_simulation.asp)

Net present value

value in today's currency of an item or service (Davenport, Enterprise Analytics, p. 22)

Network optimization

the process of striking the best possible balance between network performance and network costs, in consideration of grade of service requirements (www. yourdictionary.com)

Next best offer (NBO)

a targeted offer or proposed action for customers based on analyses of past history and behavior, other customer preferences, purchasing context, attributes of the produces, or services from which they can choose (Davenport, Enterprise Analytics, p. 83)

Nominal group technique (NGT)

a structured method for group brainstorming that encourages contributions from everyone (http://asq. org/learn-about-quality/idea-creation-tools/overview/ nominal-group.html)

Normalization

splits up data to avoid redundancy (duplication) by moving commonly repeating groups of data into new tables. Normalization therefore tends to increase the number of tables that need to be joined to perform a given query, but reduces the space required to hold the data and the number of places where it needs to be updated if the data changes (http://en.wikipedia.org/ wiki/Snowflake_schema)

Objective function

the (real-valued) function to be optimized. In a mathematical program in standard form, this is denoted f (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary. computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)

OLAP	an abbreviation for "Online Analysis and Processing"; a type of database technology that has long been used by the business community to analyze and interactively explore large financial data sets. The basic idea is that data sets are viewed as cubes with hierarchies along each axis (http://biolap.sourceforge.net/whitepaper.pdf)
OLAP cube	an array of data understood in terms of its zero or more dimensions; each cell of the cube holds a number that represents some measure of the business, such as sales, profits, expenses, budget, and forecast (http://en.wikipedia.org/wiki/OLAP_cube)
Operations management	deals with the design and management of products, processes, services, and supply chains. It considers the acquisition, development, and utilization of resources that firms need to deliver the goods and services their clients want (http://mitsloan.mit.edu/omg/om-definition.php)
Operations Research	a discipline that deals with the application of advanced analytical methods to help make better decisions (http://en.wikipedia.org/wiki/Operations_research)
Opportunity cost	the cost of an alternative that must be forgone to pursue a certain action (http://www.investopedia.com/terms/o/opportunitycost.asp)
Optimization	procedure or procedures used to make a system or design as effective or functional as possible, especially the mathematical techniques involved (http://www.thefreedictionary.com/optimization)
Pareto concept	See, 80/20 rule

Pattern recognition	in machine learning, pattern recognition is the assignment of a label to a given input value (http://en.wikipedia.org/wiki/Pattern_recognition)
Payback	the length of time required to recover the cost of an investment (http://www.investopedia.com/terms/p/paybackperiod.asp)
Pie chart	graphic depiction of data using a pie with different 'slices' to represent the relative size of different groupings of data points to the size of the whole
Precision	the degree to which repeated measurements under unchanged conditions show the same results (http://en.wikipedia.org/wiki/Accuracy_and_precision)
Predictive analytics	any approach to data mining with four attributes: an emphasis on prediction (rather than description, classification, or clustering), rapid analysis measured in hours or days (rather than the stereotypical months of traditional data mining), an emphasis on the business relevance of the resulting insights (no ivory tower analyses), and (increasingly) an emphasis on ease of use, thus making the tools accessible to business users (http://www.gartner.com/it-glossary/predictive-analytics)
Prescriptive analytics	evaluates and determines new ways of operating targeting business objective and balancing all constraints (http://www.informs.org/Community/Analytics/About-Us)

Pricing	a tactic in the simplex method, by which each variable is evaluated for its potential to improve the value of the objective function. Let $p = c_B[B^{-1}]$, where B is a basis, and c_B is a vector of costs associated with the basic variables. The vector p is sometimes called a dual solution, though it is not feasible in the dual before termination; p is also called a simplex multiplier or pricing vector. The price of the jth variable is $c_j - pA_j$. The first term is its direct cost (c_j) and the second term is an indirect cost, using the pricing vector to determine the cost of inputs and outputs in the activity's column (A_j). The net result is called the reduced cost, and its value determines whether this activity could improve the objective value (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Principal Component Analysis (PCA)	a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set (ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf)
Probability density function	the equation used to describe a continuous probability distribution (http://stattrek.com/statistics/dictionary.aspx?definition=Continuous_probability_distribution)
Problem assessment/ framing	initial step in the analytics process; involves buy in from all parties involved on what the problem is before a solution can be found
Project management	the application of knowledge, skills, and techniques to execute projects effectively and efficiently. A strategic competency for organizations, enabling them to tie project results to business goals (http://www.pmi.org/About-Us/About-Us-What-is-Project-Management.aspx)

Proprietary data   data that no other organization possesses; produced by a company to enhance its competitive posture (Davenport, Enterprise Analytics, p. 37)

Queuing theory   mathematical study of waiting in lines; results are used when making business decisions about the resources needed to provide service; research begun by A. K. Erlang (http://en.wikipedia.org/wiki/Queuing_theory on 2/20/13)

Random   of or characterizing a process of selection in which each item of a set has an equal probability of being chosen (http://dictionary.reference.com/browse/random)

Range   the difference between the maximum and minimum observations providing an estimate of the spread of the data (http://explorable.com/range-in-statistics)

Regression   a statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables) (http://www.investopedia.com/terms/r/regression.asp)

Regression analysis   statistical approach to forecasting change in a dependent variable (e.g., sales revenue) on the basis of change in one or more independent variables (e.g., population and income); AKA curve fitting or line fitting (www.businessdictionary.com)

Response surface methodology (RSM)	a surface in (n+1) dimensions that represents the variations in the expected value of a response variable (see, regression) as the values of n explanatory variables are varied. Usually the interest is in finding the combination that gives a global maximum (or minimum) (http://www.answers.com/topic/response-surface)
Return on investment (ROI)	calculations that provide a basis for comparison with other investment opportunities; typically calculated using ROI = ((Total value/benefits) – (total investment costs))/Total investment costs (Davenport; Enterprise Analytics, p. 20)
Revenue management	the science and art of enhancing revenues while selling essentially the same amount of product (http://www.ivey.uwo.ca/faculty/Peter_Bell/RM%20Ahmedabad%202005.pdf)
RFM	data related to customer relationship management; refers to recency, frequency, and monetary value of purchases (Davenport, Enterprise Analytics, p. 49)
Risk	the potential of loss (an undesirable outcome, however not necessarily so) resulting from a given action, activity, and/or inaction (http://en.wikipedia.org/wiki/Risk)

Robust optimization	a term given to an approach to deal with uncertainty, similar to the recourse model of stochastic programming, except that feasibility for all possible realizations (called scenarios) is replaced by a penalty function in the objective. As such, the approach integrates goal programming with a scenario-based description of problem data (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society. informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Scatter plot	graphic depiction of data, used to show/identify relationship between independent variables
Scenario analysis	a process of analyzing possible future events by considering alternative possible outcomes (scenarios). The analysis is designed to allow improved decision making by allowing more complete consideration of outcomes and their implications (http://www. investordictionary.com/definition/scenario-analysis#sthash.f03iNGP9.dpuf)
Scheduling	a schedule for a sequence of jobs, say j1,...,jn, is a specification of start times, say t1,...,tn, such that certain constraints are met. A schedule is sought that minimizes cost and/or some measure of time, like the overall project completion time (when the last job is finished) or the tardy time (amount by which the completion time exceeds a given deadline). There are precedence constraints, such as in the construction industry, where a wall cannot be erected until the foundation is laid (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary. computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)

Sensitivity analysis	the concern with how the solution changes if some changes are made in either the data or in some of the solution values (by fixing their value). Marginal analysis is concerned with the effects of small perturbations, maybe measurable by derivatives. Parametric analysis is concerned with larger changes in parameter values that affect the data in the mathematical program, such as a cost coefficient or resource limit (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society. informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Shadow price	an economic term to denote the rate at which the optimal value changes with respect to a change in some right-hand side that represents a resource supply or demand requirement (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Simulated annealing	an algorithm for solving hard problems, notably combinatorial programs, based on the metaphor of how annealing works: reach a minimum energy state upon cooling a substance, but not too quickly in order to avoid reaching an undesirable final state. As a heuristic search, it allows a nonimproving move to a neighbor with a probability that decreases over time. The rate of this decrease is determined by the cooling schedule, often just a parameter used in an exponential decay (in keeping with the thermodynamic metaphor). With some (mild) assumptions about the cooling schedule, this will converge in probability to a global optimum (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society. informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)

Six Sigma   a set of strategies, techniques, and tools for process improvement. It seeks to improve the quality of process outputs by identifying and removing the causes of defects (errors) and minimizing variability in manufacturing and business processes (http://en.wikipedia.org/wiki/Six_Sigma)

Spreadsheet analysis   the analysis of data using special computer software to anticipate marketing performance under a given set of circumstances (http://www.marketing-dictionary.com/s.php)

Standard deviation   measure of the unpredictability of a random variable, expressed as the average deviation of a set of data from its arithmetic mean and computed as the positive square root of the variance. Customarily represented by the lower-case Greek letter sigma (σ), it is considered the most useful and important measure of dispersion that has all the essential properties of the variance plus the advantage of being determined in the same units as those of the original data. Also called root mean square (RMS) deviation (www.businessdictionary.com)

Statistical significance probability of obtaining a test result that occurs by chance and not by systematic manipulation of data (www.businessdictionary.com)

Statistics   branch of mathematics concerned with collection, classification, analysis, and interpretation of numerical facts, for drawing inferences on the basis of their quantifiable likelihood (probability). Statistics can interpret aggregates of data too large to be intelligible by ordinary observation because such data (unlike individual quantities) tend to behave in regular, predictable manner. It is subdivided into descriptive and inferential statistics (www.businessdictionary.com)

Stepwise regression	a semi-automated process of building a model by successively adding or removing variables based solely on the t-statistics of their estimated coefficients (http://people.duke.edu/~rnau/regstep.htm)
Supply chain management	the active management of supply chain activities to maximize customer value and achieve a sustainable competitive advantage (http://scm.ncsu.edu/scm-articles/article/what-is-supply-chain-management)
System dynamics	a computer-aided approach to policy analysis and design. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems (http://www.systemdynamics.org/what_is_system_dynamics.html)
Tolerance	an approach to sensitivity analysis in linear programming that expresses the common range that parameters can change while preserving the character of the solution (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Traveling salesman problem (TSP)	given n points and a cost matrix $[c_{ij}]$, a tour is a permutation of the n points. The points can be cities, and the permutation the visitation of each city exactly once, then returning to the first city (called home). (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)
Uncertainty	the estimated amount or percentage by which an observed or calculated value may differ from the true value (http://www.thefreedictionary.com/uncertainty)

Validation (of a model)	determining how well the model depicts the real-world situation it is describing (http://www.easterbrook.ca/steve/2010/11/the-difference-between-verification-and-validation/)
Variability	describes how spread out or closely clustered a set of data is (http://en.wikipedia.org/wiki/Variability)
Variable cost	a periodic cost that varies in step with the output or the sales revenue of a company. Variable costs include raw material, energy usage, labor, distribution costs, etc. (http://www.businessdictionary.com/definition/variable-cost.html)
Variance	a parameter in a distribution that describes how far the values are spread apart. Variance is a characteristic of some probability distribution, which distinguishes the concept of variance from the ways to estimate it from sample data(http://en.wikipedia.org/wiki/Variance)
Variation reduction	reference to process variation where reduction leads to stable and predication process results (http://www.businessdictionary.com)
Vehicle routing problem (VRP)	finding optimal delivery routes from one or more depots to a set of geographically scattered points (e.g., population centers). A simple case is finding a route for snow removal, garbage collection, or street sweeping (without complications, this is akin to a shortest path problem). In its most complex form, the VRP is a generalization of the TSP, as it can include additional time and capacity constraints, precedence constraints, and more (A. Holder, editor. Mathematical Programming Glossary. INFORMS Computing Society, http://glossary.computing.society.informs.org/, 2006-08. Originally authored by Harvey J. Greenberg, 1999-2006.)

Verification
(of a model)

includes all the activities associated with the producing high quality software: testing, inspection, design analysis, specification analysis (http://www.easterbrook. ca/steve/2010/11/the-difference-between-verification-and-validation/)

Web analytics

ability to use data generated through Internet-based activities; typically used to assess customer behaviors; see also, RFM (Davenport, Enterprise Analytics, p. 49-51)

Yield

percentage of 'good' product in a batch; has three main components: functional (defect driven), parametric (performance driven), and production efficiency/ equipment utilization (http://www-inst.eecs.berkeley. edu/~ee290h/fa05/Lectures/PDF/lecture%201%20 intro%20IC%20Yield.pdf)

These questions will never be on the CAP® certification exam: they are here solely as study aids. All questions on the certification exam are multiple choice with four possible correct answers of which only one is correct.

1. What are the 5 W's?

2. What is a stakeholder?

3. How could a problem not be amenable to an analytics solution?

4. Suppose that the business problem is that the organization wants to increase sales by increasing cross-selling to existing customers. Your project sponsor looks to you to tell her how the organization can get there based on the data at hand. What's your first move?

    a. Dive into existing customer interaction data

    b. Ask your sponsor if she has a particular customer segment in mind

    c. Talk with marketing to see what they have planned for the next sales campaign

    d. Ask your sponsor what the actual numeric target of increased sales is overall

5. Your sponsor has come back with a numeric goal of increasing sales from an average of $10,000 per customer to $11,000 per customer in the next 12 months, what's your next move?

    a. See what price/sales volume data exist to see if the organization's prices match value

    b. See what sales by customer data exist

    c. Create hypotheses of which customer segments could be cross-sold

    d. Explore whether there are any other related business goals

6. You now have a little more information from the project sponsor, along with several rumors from other sources. You know that you should base the cost of increased sales over current levels at the marginal cost, rather than the fully allocated cost; that the company has to maintain at least the same return on sales as it currently has as the sales increase from 10,000 per customer to 11,000 per customer; and that top-line revenue must also increase by 10% (i.e., you can't get there by dropping your lowest performing customers). Once you've listed these assumptions or rules in your project charter, what's next?

   a. Start creating your input/output diagrams about what drives current customers to buy more

   b. Talk with your marketing and data groups to see what data exist

   c. Figure out how the increased sales goal should be broken down into metrics

   d. Run a conjoint analysis to see if existing products can be tweaked to be worth more money

7. Speaking of reviews, which of these groups should <u>NOT be invited?</u>

   a. Data group

   b. Sales & Marketing

   c. Manufacturing

   d. Contracts

8. Describe the main differences between discrete-event simulation and Monte Carlo simulation.

9. A post office area manager received many complaints that the only branch she has in the north side of the town has a very long waiting time. She hired you as a consultant to recommend justifying opening new positions in her branch. What would be a relevant methodology to use?

   a. Monte Carlo simulation

   b. Queuing theory

   c. Data mining

   d. Linear programming

10. A major aircraft manufacturing company is intending to determine the main causes for fatal failures in their battery system. The best methodology to use to pin point the root causes is:

    a. conduct a well-prepared design of experiments.

    b. use historical data to relate failures to potential causes.

    c. Simulate the process with all the failure modes

    d. Choice B or C

11. In mapping different X's to a Y, the advantage of using linear regression to a backpropagation artificial neural network (ANN) is:

    a. regression is more accurate in predicting Y's given X's compared to ANN.

    b. regression can handle more variables than ANN.

    c. regression handles data in a visible and transparent manner compared to ANN, which is perceived to be a black-box methodology.

    d. regression is more able to handle outliers.

12. You are given three months to solve an analytics problem and the needed data will require two months to collect. What would be the strategy with the best outcome?

    a. Wait until the data are available to choose the best methodology

    b. Refuse to work on this project

    c. Ignore the data and design a tool that fits all possible scenarios

    d. Start developing the model with a template containing approximate numbers

13. One good methodology to reduce the dimensionality of a set of data is to use:

    a. principal component analysis (PCA).

    b. linear programming.

    c. discrete-event simulation.

    d. artificial intelligence.

14. You are given a set of data to be utilized for a model. Their level of accuracy is within +/- 20%. What approach and/or software would you use for the problem?

    a. Approach and/or software that deals with data at +/- 1% accuracy level

    b. Approach and/or software that deals with data at +/- 0.01% accuracy level

    c. Approach and/or software that deals with data at +/- 10% accuracy level

    d. Approach and/or software that deals with data at +/- 30% accuracy level

15. You are asked to establish a model to map many independent variables (X's) to one dependent variable (Y). The model should explain the level of significance of the X's to Y and their level of correlation. What is the first methodology to come to mind in this situation?

    a. Stepwise regression

    b. Fuzzy logic

    c. Artificial neural network

    d. Monte Carlo simulation

16. In INFORMS CAP® study guide, models are classified as:

    a. prescriptive, simulation, and predictive.

    b. descriptive, prescriptive, and predictive.

    c. analytical, soft skills, and descriptive.

    d. simulation, optimization, data mining, and statistics.

17. A factory has skilled workers that operate complicated equipment and there is a need to transfer the knowledge to new hires. The procedure cannot be explained in a crisp manner with exact numbers. For example, the operator cannot explain what the right temperature and pressure are to maximize the strength of the material at a certain condition. They simply just know by experience. One good candidate approach to model that variables and rules is:

   a. fuzzy logic.

   b. neural network.

   c. linear regression.

   d. logistic regression.

18. Visualization is more closely related to which of the following analytics methodology categories?

   a. Prescriptive

   b. Descriptive

   c. Soft skills

   d. Predictive

19. A proper methodology to handle missing data is:

   a. principal component analysis.

   b. stepwise regression.

   c. decision tree.

   d. Markov chain.

20. A chemical plant is under study to identify the bottleneck in its operation to facilitate scheduling. One proper methodology to model the plant is:

   a. system dynamics.

   b. discrete-event simulation.

   c. Markov chain.

   d. fuzzy logic.

21. You are given a problem by a client in which you need to determine the right amount to be purchased from what location so the total cost of manufacturing, transportation, and duties is minimized. The first methodology to come in mind to model this problem is:

    a. step-wise regression.

    b. mixed-integer programming.

    c. linear programming.

    d. logistic regression.

22. Genetic algorithm, Tabu search, and ant colony optimization are examples of optimization algorithms that are inspired by natural phenomenon and are examples of the following type of analytics methodologies:

    a. Metaheuristics

    b. Simulation

    c. Pattern recognition

    d. Visualization

23. Once you've built your model how do you know that the model will still answer your business problem?

24. In the business problem framing chapter, there's an example of a manufacturing plant that has poor on-time performance. Imagine that you've built a simulation model of the plant that shows that it should be able to achieve much better results without requiring any new investment. What concerns might your stakeholders have?

25. When should you retire a model?

    a. When its replacement has been validated

    b. When a change in business conditions invalidate its assumptions

    c. Both a and b

    d. Neither a nor b

26. How often should model maintenance be done?

    a. When underlying assumptions change

    b. When it is ported to a new system

    c. When the data it uses changes its format

    d. When it is transferred to a new owner

27. What will happen if you don't ever bother to evaluate model performance and returns over time?

28. Which of the following BEST describes the data and information flow within an organization?

    a. Information assurance

    b. Information strategy

    c. Information mapping

    d. Information architecture

29. A multiple linear regression was built to try to predict customer expenditures based on 200 independent variables (behavioral and demographic). 10,000 randomly selected rows of data were fed into a stepwise regression, each row representing one customer. 1,000 customers were male, and 9,000 customers were female. The final model had an adjusted R-squared of 0.27 and seven independent variables. Increasing the number of randomly selected rows of data to 100,000 and rerunning the stepwise regression will MOST likely:

    a. have negligible impact upon the adjusted R-squared.

    b. increase the impact of the male customers.

    c. change the heteroskedasticity of the residuals in a favorable manner.

    d. decrease the number of independent variables in the final model.

30. A clothing company wants to use analytics to decide which customers to send a promotional catalogue in order to attain a targeted response rate. Which of the following techniques would be the MOST appropriate to use for making this decision?

    a. Integer programming

    b. Logistic regression

    c. Analysis of variance

    d. Linear regression

31. Which of the following is an effective optimization method?

    a. Analysis of variance (ANOVA)

    b. Generalized linear model (GLM)

    c. Box-Jenkins Method (ARIMA)

    d. Mixed integer programming (MIP)

32. A box and whisker plot for a dataset will MOST clearly show:

    a. the difference between the 50th percentile and the median.

    b. the 90% confidence interval around the mean.

    c. where the [actual-predicted] error value is not zero.

    d. if the data is skewed and, if so, in which direction.

33. In the initial project meeting with a client for a new project, which of the following is the MOST important information to obtain?

    a. Timeline and implementation plan

    b. Analytical model to use

    c. Business issue and project goal

    d. Available budget

34. Which of the following statements is true of modeling a multi-server checkout line?

    a. A queuing model can be used to estimate service rates.

    b. A queuing model can be used to estimate average arrivals.

    c. Variability in arrival and service times will tend to play a critical role in congestion.

    d. Poisson distributions are not relevant.

35. A company is considering designing a new automobile. Their options are a design based on current gasoline engine technology or a government proposed «Green» technology. You are a government official whose job is to encourage automakers to adopt the «Green» technology. You cannot provide funding for development costs, but you can provide a subsidy for every car sold. The development costs and the wholesale price, in thousands of dollars, of the cars are shown in the table below:

	GASOLINE TECHNOLOGY (NUMBERS IN $ THOUSANDS)	"GREEN" TECHNOLOGY (NUMBERS IN $ THOUSANDS)
WHOLESALE PRICE/VEHICLE	25	40
VARIABLE COST/VEHICLE	15	35
FIXED COST	100,000	200,000

How large a subsidy per vehicle sold will be required, assuming there will be enough demand to motivate the switch?

    a. Greater than $5000

    b. Less than $5000

    c. Cannot be determined

    d. Equal to $5000

36. A furniture maker would like to determine the most profitable mix of items to produce. There are well-known budgetary constraints. Each piece of furniture is made of a predetermined amount of material with known costs, and demand is known. Which of the following analytical techniques is the MOST appropriate one to solve this problem?

    a. Optimization

    b. Multiple regression

    c. Data mining

    d. Forecasting

37. You have simulated the Net Present Value (NPV) of a decision. It ranges between –$10 million and +$10 million. To best present the likelihood of possible outcomes, you should:

    a. present a single NPV estimate to avoid confusion.

    b. present a histogram to show likelihood of various NPVs.

    c. trim all outliers to present the most balanced diagram.

    d. relax constraints associated with extreme points in the simulation.

38. A company ships products from a single dock at their warehouse. The time to load shipments depends on the experience of the crew, products being shipped and weather. The company thinks there is significant unmet demand for their products and would like to build another dock in order to meet this demand. They ask you to build a model and determine if the revenue from the additional products sold will cover the cost of the second dock within two years of it becoming operational. Which of the following is the MOST appropriate modeling approach and justification?

    a. Optimization because it is a transportation problem.

    b. Optimization because the company's objective is to maximize profit and because capacity at the dock is a limited resource.

    c. Forecasting because you can determine the throughput at the dock, calculate the net revenue and compare this with the cost of the new dock.

    d. Discrete event simulation because there are a sequence of random events through time.

39. Two investors who have the same information about the stock market buy an equal number of shares of a stock. Which of the following statements must be true?

   a. The risks for the two investors are statistically independent.

   b. Both investors are subject to the same risks.

   c. Both investors are subject to the same uncertainty.

   d. If the investors are optimistic, they should have borrowed rather than bought the shares.

40. A project seeks to build a predictive data-mining model of customer profitability based upon a set of independent variables including customer transaction history, demographics, and externally purchased credit-scoring information. There are currently 100,000 unique customers available for use in building the predictive model. Which of the following strategies would reflect the BEST allocation of these 100,000 customer data points?

   a. Use 70,000 randomly selected data points when building the model, and hold the remaining 30,000 out as a test dataset.

   b. Use all 100,000 data points when building the model.

   c. Randomly partition the data into 4 datasets of equal size, build four models and take their average.

   d. Use 1,000 randomly selected data points when building the model.

41. Conjoint analysis in market research applications can:

   a. give its best estimates of customer preference structure based on in-depth interviews with a small number of carefully chosen subjects.

   b. only trade off relative importance to customers of features with similar scales.

   c. allow calculation of relative importance of varying features and attributes to customers.

   d. only trade off among a limited number of attributes and levels.

42. One of the main advantages of tree-based models and neural networks is that they:

    a. are easy to interpret, use, and explain.

    b. build models with higher R-squared than other regression techniques.

    c. reveal interactions without having to explicitly build them into the model.

    d. can be modeled even when there is a significant amount of missing data.

43. The monthly profit made by a clothing manufacturer is proportional to the monthly demand, up to a maximum demand of 1000 units, which corresponds to the plant producing at full capacity. (Any excess demand over 1000 units will be satisfied by some other manufacturer, and hence yield no additional profit.) The monthly demand is uncertain, but the average demand is reliably estimated at 1000 units. At this level of demand the monthly profit is $3,000,000. Which of the following statements must be true of the expected monthly profit, P?

    a. P can have any positive value.

    b. P is possibly greater than $3,000,000.

    c. P is equal to $3,000,000.

    d. P is less than $3,000,000.

44. After building a predictive model and testing it on new data, an under prediction by a forecasting system can be detected by its:

    a. negative-squared.

    b. bias.

    c. mean absolute deviation.

    d. mean squared error.

45. All times in the decision tree below are given in hours. What is the expected travel time (in hours) of the optimal (minimum travel time) decision?
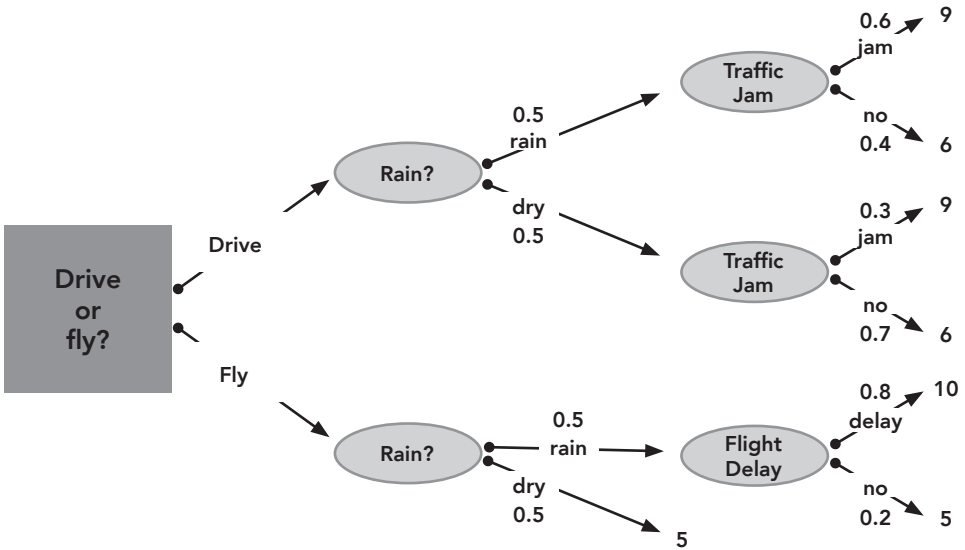

a. 7.8

b. 6.9

c. 7.4

d. 7.0

46. An analytics professional is responsible for maintaining a simulation model that is used to determine the staffing levels required for a specific operational business process. Assuming that the operational team always uses the number of staff determined by the model, which of the following is the MOST important maintenance activity?

a. Ensure that all the model input data items are available when needed.

b. Determine if there has been a change in model accuracy over time.

c. Ensure that all users are reviewing the model results in a timely fashion.

d. Determine that the model's reports are understood by the users.

47. A segmentation of customers who shop at a retail store may be performed using which of the following methods?

    a.   Monte Carlo Markov Chain and ANOVA

    b.   Clustering, factor and control charts

    c.   Decision tree and recursive function analyses

    d.   Clustering and decision trees

48. In the diagram below, what is true of Strategy B compared to Strategy A?
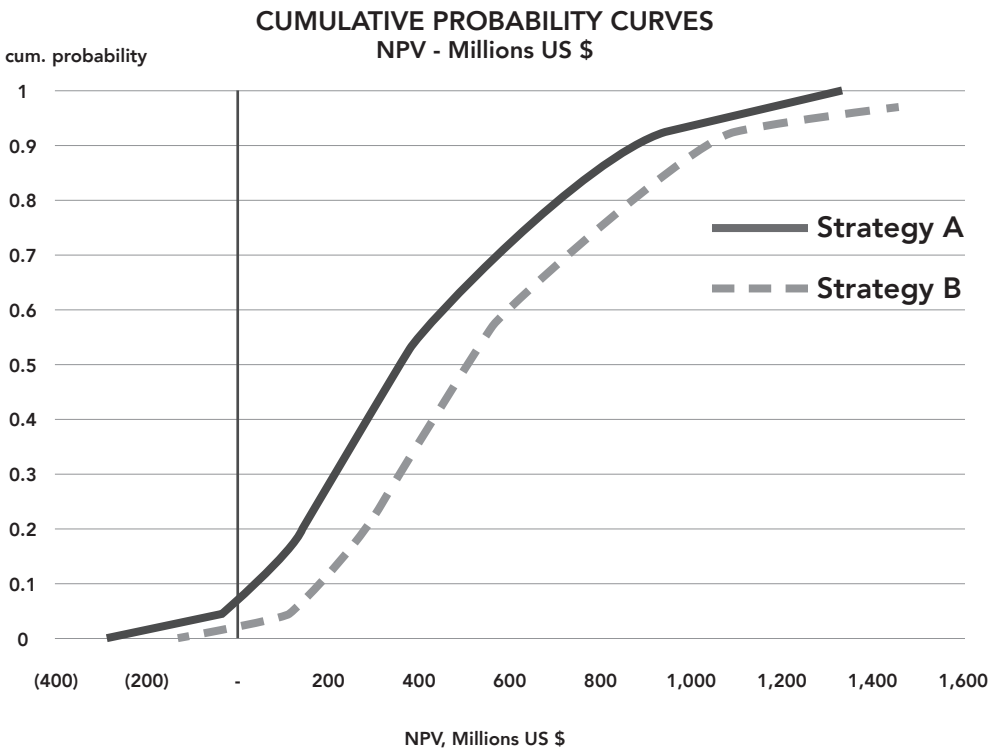
CUMULATIVE PROBABILITY CURVES
NPV - Millions US $

    a.   Strategy B exhibits stochastic (probabilistic) dominance over Strategy A.

    b.   Strategy B has the same downside risk as Strategy A since the curves have the same shape.

    c.   Strategy B must have the same uncertainties impacting it as Strategy A because the curves are so similar in shape.

    d.   Strategy A exhibits stochastic (probabilistic) dominance over strategy B.

49. Each month you generate a list of marketing leads for direct mail campaigns. Which of the following should you do before the list is used?

    a. Exclude people who were on the list the previous month.

    b. Retain x% of the leads as control for performance measurement.

    c. Remove opt-outs.

    d. Exclude people who were never on the list.

50. When analyzing responses of a survey of why people like a certain restaurant, factor analysis could reduce the dimension in which of the following ways?

    a. Collapse several survey questions regarding food taste, health value, ingredients and consistency into one general unobserved "food quality" variable.

    b. Condense similar survey respondent answers into clusters of like-minded customers for market segment analysis.

    c. Reduce the variability of individual subject ratings by centering each respondent's ratings around his or her average rating.

    d. Decrease variability by analyzing inter-rater reliability on the question items before offering the survey to a wide number of respondents.

51. A preferred method or best practice for organizing data in a data warehouse for reporting and analysis is:

    a. transactional-based modeling.

    b. multidimensional modeling.

    c. relation-based modeling.

    d. tuple-based modeling.

1. Who are the stakeholders,

    What is the problem,

    Where is the problem occurring,

    When does the problem occur, and

    Why does the problem occur?

2. Stakeholders are all who are affected by the problem and its solution. Note that this may include more than those in the initial meetings and those in charge of the problem solution.

3. Problems may be constrained by limitations of the tools, methods, and data available or the feasibility of solution.

4. Note that your sponsor didn't give you much information to go on, and you don't know what your goal really is, except that you know you're looking to get more sales per customer. There's not enough to go on here to start to formulate the problem. Choice D would be the best response to start to get some numbers to go with the business' goal.

5. Even given the statement above, you don't yet have a complete view of the business problem. You don't know why the organization has chosen to focus its attention on increasing sales per customer. Without that, you don't know what margins are acceptable on those sales. You may assume that general business rules apply and that you should assume that any sales under a 20% margin are inherently unprofitable and should be rejected. But without surfacing and clarifying that assumption and many others, you don't know if it is valid or not. You have to ask and keep asking until you know what assumptions are valid. Again, choice D is the most appropriate answer.

6. Here the most appropriate answer is choice A. This is important because if you go straight to looking at data, your hypotheses about what's important will be inherently biased by the existing data and explanations. If the answer were in your existing explanations, you probably wouldn't have the problem in the first place. But now that you have the initial set of drivers, you can start talking with your data group and decomposing your metrics to allocate the increased performance to performing groups. Any group with changing goals needs to be on your stakeholder list and part of the reviews.

7. Any group with changing requirements needs to be invited. If you plan on selling more items, then the manufacturing group needs to be part of the discussion so they can advise on how much they can actually produce before requiring more investment for another line, more employees, etc.

8. Monte Carlo simulation is about generating random numbers and processes them to predict another variable without focusing necessarily on the accumulated queues and the impact of time. On the other hand, the focus of discrete event simulation is to study the accumulated queues as time goes. A discrete event simulation may include Monte Carlo runs as we can run random numbers in DES, but not necessarily.

9. b. Queuing theory

10. d. Choice B or C

11. c. regression handles data in a visible and transparent manner compared to ANN, which is perceived to be a black-box methodology.

12. d. Start developing the model with a template containing approximate numbers

13. a. principal component analysis (PCA).

14. c. Approach and/or software that deals with data at +/- 10% accuracy level

15. a. Stepwise regression

16. b. descriptive, prescriptive, and predictive.

17. a. fuzzy logic.

18. b. Descriptive

19. a. principal component analysis.

20. b. discrete-event simulation.

21. b. mixed-integer programming.

22. a. Metaheuristics

23. The answer is to go back to the original question or problem and see if that has been answered. There may be times when the original question or problem may have become only a part of the solution, but it still needs to have been answered.

24. Among other things, stakeholders may be concerned with the implications of the solution, the future impact on their business, whether the new solution will lead to more on time performance in the long run, the ease of implementation, impact on personnel of changes in processes, and other concerns related to their way of doing business.

25. c. Both a and b. If a change in business conditions has occurred that invalidate the assumptions of the original model, a new or revised model should be fielded and tested and validated before being deployed as a replacement.

26. a. While maintenance is continual over the life of a model, maintenance is required when the underlying assumptions change.

27. If the model performance is not evaluated, over time the returns may become skewed and may not provide accurate answers to the original question.

28. d. Information architecture refers to the analysis and design of the data stored by information systems, concentrating on entities, their attributes, and their interrelationships. It refers to the modeling of data for an individual database and to the corporate data models that an enterprise uses to coordinate the definition of data in several (perhaps scores or hundreds) distinct databases.

29. a. have no impact upon the adjusted R-squared. The increase in size of the data will not impact the adjusted R-squared calculation because both samples are sufficiently large randomly selected subsets of data.

30. b. Logistic regression This type of classification model is often used to predict the outcome of a categorical dependent variable (response vs. no response) based on one or more predictor variables, so this is the most appropriate answer. The goal of the analytics in the stated problem is to determine who is most likely to respond, and the binary nature of this predicted outcome is provided by logistic regression.

31. d. Mixed integer programming (MIP) This is a mathematical optimization technique used when one or more of the variables are restricted to be integers. It is an effective optimization model.

32. d. if the data is skewed and, if so, in which direction. A box and whisker plot, sometimes just called a "box plot," was invented by John Tukey as a way to graphically display the distribution of data. The ends of the box are at the first and third quartiles, and there is a line somewhere in the box representing the median value. The whiskers extend either to the minimum and maximum values in the data set, or possibly less if they do not include points identified as outliers.

33. c. Business issue and project goal. Understanding the business issue and project goal provides a sound foundation on which to base the project.

34. c. Variability in arrival and service times will tend to play a critical role in congestion. Arrival and service time distributions are inputs to a queuing model that would be used to model a checkout line and directly influence congestion.

35. a. Greater than $5000

    If we consider the profit from an individual vehicle to be the wholesale price minus the variable cost, we see that the profit from a Gasoline Technology vehicle is $25K–$15K = $10K. Similarly, the profit from a "Green" Technology vehicle is $40K–$35K = $5K. In order to make up for this difference in lost profit, the subsidy provided to the automaker would have to be at least $5K (the difference between $10K and $5K). In addition, the subsidy would need to be greater than $5000 so that the auto makers would be able to recover their increased fixed costs at a reasonable level of demand.

36. a. Optimization The problem statement describes an optimization problem: the furniture maker's objective function is to maximize his profit. The decision variables are the amount of each item to produce, and the constraints are that he must meet demand and be within his budget. Optimization is the most appropriate technique to solve this problem.

37. b. present a histogram to show likelihood of various NPVs. Net Present Value (NPV) takes as input a time series of cash flow (both incoming and outgoing) and a discount rate and outputs a price. By showing a histogram (a graphical representation of the distribution of data), it is possible to see how likely various NPVs (beyond the given minimum and maximum) are to occur. This would be useful information to have when considering a decision, especially since the range of outcomes includes $0, meaning the decision could result in a profit or a loss.

38. d. Discrete event simulation because there are a sequence of random events through time. The time to load shipments depends on the experience of the crew, products being shipped and weather. Given there is a sequence of random events through time, discrete event simulation is the most appropriate modeling approach.

39. c. Both investors are subject to the same uncertainty regarding the stock market.

40. a. Use 70,000 randomly selected data points when building the model, and hold the remaining 30,000 out as a test dataset. This split provides sufficient data to build the model and sufficient data to test the model. This is the best allocation of the customer data points. (A common 'rule of thumb' is to use about two thirds of the data to build the model and one third to test it)

41. c. allow calculation of relative importance of varying features and attributes to customers. Conjoint analysis by definition maps consumer preference structures into mathematical tradeoffs, and was designed to allow a marketer to compare the relative utility of varying features and attributes.

42. c. reveal interactions without having to explicitly build them into the model. Tree-based models and neural networks are employed to find patterns in the data that were not previously identified (or input into the model building process).

43. d. P is less than $3,000,000. When the demand is 1000 or greater, the profit is $3,000,000. But when the demand is less than 1000, the profit is less than $3,000,000. Given this and that the average demand is 1000 units, the expected monthly profit must be less than $3,000,000.

44. b. bias. The bias measures the difference, including the direction of the estimate and the right answer. Depending on whether it's positive or negative, it will show whether there is an over or under estimate.

45. d. 7.0 To answer this question, one needs to solve the decision tree using the "roll back" technique. Continuing back the bottom branch of the tree, the expected time if you fly is $(0.5)(9.0) + (0.5)(5) = 7.0$ hours. Now, when faced with the "drive or fly" decision, you should choose to fly (since 7.0 hours is less than 7.35 hours). Thus, answer d) 7.0 hours is the expected travel time of the optimal (or minimal travel time) decision.

46. b. Determine if there has been a change in model accuracy over time. The most important maintenance activity for the analytics professional responsible for maintaining the simulation model is to monitor the accuracy of the model over time. If there has been a change in accuracy, the analytics professional may need to revisit the assumptions of the model.

47. d. Clustering and decision trees Customer segmentation consists of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, e.g., age, gender, interests, spending habits and so on. The purpose of customer segmentation is to allow a company to target specific groups of customers effectively and allocate marketing resources to best effect. Two ways to do this segmentation are clustering and decision trees

48. a. Strategy B exhibits stochastic (probabilistic) dominance over Strategy A.

Because the cumulative probability curve for Strategy B is below (or to the right) of the corresponding curve for Strategy A, it can be said that Strategy B exhibits stochastic dominance (SD) over Strategy A. B stochastically dominates A when, for any good outcome x, B gives at least as high a probability of receiving at least x as does A, and for some x, B gives a higher probability of receiving at least x. Since the curves do not cross, B stochastically dominates A.

49. c. Remove opt-outs. The list of marketing leads should not include people or organizations that have opted out.

50. a. Collapse several survey questions regarding food taste, health value, ingredients and consistency into one general unobserved "food quality" variable. Factor analysis is a statistical method used to describe variability among observed variables in terms of a potentially lower number of unobserved variables called factors. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset.

51. b. multidimensional modeling.

Multidimensional modeling is the optimum way to organize data in a data warehouse for analysis. It is associated with OLAP (On-line Analytical Processing). OLAP data is organized in cubes that can be taken directly from the data warehouse for analysis.

For more information on the review questions numbered 28 to 51, see https:// www.certifiedanalytics.org

Domain I — Business Problem (Question) Framing

Davenport T, Kim J (2013) *Keeping up with the Quants: Your Guide to Understanding and Using Analytics* (Harvard Business Review Press, Boston).

Framing the problem at https://www.boundless.com/business/management/decision-making/observation-framing-the-problem/.

Kirkwood CW (1997) *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets* (Duxbury Press, Pacific Grove, CA).

Lindstrom C (2009) How to write a problem statement, March 18, http://www.ceptara.com/blog/how-to-write-problem-statement.

Nixon NW (2013) Focus first on framing, not solving, the problem, April 18, http://philadelphia.regionsbusiness.com/print-edition-commentary/focus-first-on-framing-not-solving-the-problem/.

Seelig T (2013) Shift your lens: The power of re-framing problems. Seelig T, ed. *inGenius: A Crash Course on Creativity* (HarperOne, New York), http://stvp.stanford.edu/blog/?p=6435.

Spradlin D (2012) The power of defining the problem, September 25, http://blogs.hbr.org/cs/2012/09/the_power_of_defining_the_prob.html.

Domain II — Analytics Problem Framing

Albright SC, Winston W, Zappe C (2011) *Data Analysis and Decision Making*, 4th ed. (South-Western Cengage Learning, Mason, OH).

Covey S (2004) *The 7 Habits of Highly Effective People* (Simon & Schuster, New York).

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. Science 185(4157):1124–1131.

Domain III — Data

Hubbard DW (2010) *How to Measure Anything: Finding the Value of "Intangibles" in Business*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Hillier F, Hillier M (2013) *Introduction to Management Science: A Modeling and Case Study Approach*, 5th ed. (McGraw-Hill Higher Education, New York).

Vose D (2008) *Risk Analysis: A Quantitative Guide*, 3rd ed. (John Wiley & Sons, Chichester, UK).

Domain IV — Methodology (Approach) Selection

Big data: The next frontier for innovation, competition, and productivity, a McKinsey & Company report. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Neter J, Kutner M, Nachtsheim C, Wasserman W (1996) *Applied Linear Statistical Models*, 4th ed. (McGraw-Hill/Irwin, New York).

Domain V — Model Building & Domain VII – Model Life Cycle Management

Berry MJA, Linoff GS (1999) *Mastering Data Mining: The Art and Science of Customer Relationship Management* (Wiley, New York).

Clemen RT (1997) *Making Hard Decisions: An Introduction to Decision*, 2nd ed. (Duxbury Press, Pacific Grove, CA).

Few S (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, 2nd ed. (Analytics Press, Burlingame, CA).

Hand DJ, Mannila H, Smyth P (2001) *Principles of Data Mining* (MIT Press, Boston).

Hillier FS, Lieberman GJ (2005) *Introduction to Operations Research*, 8th ed. (McGraw-Hill, New York).

Law AM, Kelton DW (2000) *Simulation Modeling and Analysis*, 3rd ed. (McGraw-Hill, New York).

Ross SM (2010) *Introductory Statistics*, 3rd ed. (Academic Press, Burlington, MA).

Siegel E (2013) *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (Wiley, New York).

Tufte ER (2001) *The Visual Display of Quantitative Information*, 2nd ed. (Graphics Press, Cheshire, CT).

Domain VI — Deployment

Chapman P, et al., CRISP-DM 1.0 Step by Step data mining guide, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf and http://www.the-modeling-agency.com/crisp-dm.pdf.

Laursen GHN, Thorlund J (2010) *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting* (John Wiley & Sons, Hoboken, NJ).

Domain VII — Lifecycle Maintenance

Chapman P, et al., CRISP-DM 1.0 Step by Step data mining guide, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf and http://www.the-modeling-agency.com/crisp-dm.pdf.

Wirth R (2000) CRISP-DM: Towards a standard process model for data mining. *Proc. Fourth Internat. Conf. Practical Appl. Knowledge Discovery Data Mining*, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133.

FURTHER READING

Albright SC, Winston W, Zappe C (2011) *Data Analysis and Decision Making*, 4th ed. (South-Western Cengage Learning, Mason, OH).

Bartlett R (2013) *A Practitioner's Guide to Business Analytics: Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy* (McGraw-Hill, New York).

Bennett G, Levis J (2013) Steering toward analytics certification. *OR/MS Today* (June).

Berry MJA, Linoff GS (1999) *Mastering Data Mining: The Art and Science of Customer Relationship Management* (Wiley, New York).

Big data: The next frontier for innovation, competition, and productivity. Report, McKinsey & Company. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Breeden J (2013) *Tipping Sacred Cows: Kick the Bad Work Habits that Masquerade as Virtues* (Jossey-Bass, San Francisco, CA).

Brohaugh W (2007) *Write Tight: Say Exactly What You Mean With Precision and Power* (Sourcebooks, Naperville, IL).

Chapman P, et al., CRISP-DM 1.0 Step by Step data mining guide, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf and http://www.the-modeling-agency.com/crisp-dm.pdf.

Clemen RT (1997) *Making Hard Decisions: An Introduction to Decision,* 2nd ed. (Duxbury Press, Pacific Grove, CA).

Covey S (2004) *The 7 Habits of Highly Effective People* (Simon & Schuster, New York).

Davenport T, Harris J (2110) *Analytics at Work: Smarter Decision, Better Results* (Harvard Business Review Press, Boston).

Davenport T, Kim J (2013) *Keeping up with the Quants: Your Guide to Understanding and Using Analytics* (Harvard Business Review Press, Boston).

Duarter N (2012) *HBR Guide to Persuasive Presentations* (Harvard Business Review Press, Boston).

Eckerson W (2012) *Secrets of Analytical Leaders: Insights from Information Insiders* (Technics Publications, Westfield, NJ).

Few S (2012) *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, 2nd ed. (Analytics Press, Burlingame, CA).

Framing the problem at https://www.boundless.com/business/management/decision-making/observation-framing-the-problem/.

Franks B (2012) *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics* (John Wiley & Sons, Hoboken, NJ).

Hand DJ, Mannila H, Smyth P (2001) *Principles of Data Mining* (MIT Press, Boston).

Hillier F, Hillier M (2013) *Introduction to Management Science: A Modeling and Case Study Approach*, 5th ed. (McGraw-Hill Higher Education, New York).

Hillier FS, Lieberman GJ (2010) *Introduction to Operations Research*, 9th ed. (McGraw-Hill, New York).

Hubbard DW (2010) *How to Measure Anything: Finding the Value of "Intangibles" in Business*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Jarman K (2013) *The Art of Data Analysis: How to Answer Almost Any Question Using Basic Statistics* (John Wiley & Sons, Hoboken, NJ).

Kirkwood CW (1997) *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets* (Duxbury Press, Pacific Grove, CA).

The Ladder of Inference: Avoiding "Jumping to Conclusions", http://www.mindtools.com/pages/article/newTMC_91.htm.

Law AM, Kelton DW (2006) *Simulation Modeling and Analysis*, 4th ed. (McGraw-Hill, New York).

Laursen GHN, Thorlund J (2010) *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting* (John Wiley & Sons, Hoboken, NJ).

Lindstrom C (2009) *How to write a problem statement*, March 18, http://www.ceptara.com/blog/how-to-write-problem-statement.

Mayer Schonberger V, Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, New York).

Neter J, Kutner M, Nachtsheim C, Wasserman W (1996) *Applied Linear Statistical Models*, 4th ed. (McGraw-Hill/Irwin, New York).

Nixon NW (2013) *Focus first on framing, not solving, the problem*, April 18, http://philadelphia.regionsbusiness.com/print-edition-commentary/focus-first-on-framing-not-solving-the-problem/.

Philllips J (2013) *Building a Digital Analytics Organization: Creating Value by Integrating Analytical Processes, Technology, and People into Business Operations* (Pearson, Upper Saddle River, NJ).

Pink D (2013) *To Sell is Human: The Surprising Truth about Moving Others* (Riverhead Books, New York).

Provost F, Fawcett T (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking* (O'Reilly Media, Sebastopol, CA).

Redman T (2001) *Data Quality: The Field Guide* (Digital Press, Woburn, MA).

Ross SM (2010) *Introductory Statistics*, 3rd ed. (Academic Press, Burlington, MA).

Sashihara S (2011) *The Optimization Edge: Reinventing Decision Making to Maximize All Your Company's Assets* (McGraw-Hill, New York).

Savage S (2012) *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty* (John Wiley & Sons, Hoboken, NJ).

Saxena R, Srinivasan A (2012) *Business Analytics: A practitioner's Guide.* (Springer, New York).

Seelig T (2013) Shift your lens: The power of re-framing problems. Seelig T, ed. *inGenius: A Crash Course on Creativity* (HarperOne, New York), http://stvp.stanford.edu/blog/?p=6435.

Shmueli G (2012) *Practical Time Series Forecasting: A Hands-On Guide* (Springer, New York).

Siegel E (2013) *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (Wiley, New York).

Silver N (2012) *The Signal and the Noise: Why Most Predictions Fail but Some Don't* (Penguin Press, New York).

Soares S (2013) *Big Data Governance: An Emerging Imperative* (MC Press Online, Boise, ID).

Spitzer DR (2007) *Transforming Performance Management: Rethinking the Way We Measure and Drive Organizational Success.* (AMACOM, New York).

Spradlin D (2012) The power of defining the problem, September 25, http://blogs.hbr.org/cs/2012/09/the_power_of_defining_the_prob.html.

Taylor J (2011) *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics* (Pearson Education, Boston, MA).

Timmer J (2013) Applying science to communicate science: Right now, it's hard to find relevant information on how to do it well, August 1, http://arstechnica.com/staff/2013/08/applying-science-to-communicate-science/.

Tufte ER (2001) *The Visual Display of Quantitative Information*, 2nd ed. (Graphics Press, Cheshire, CT).

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Vose D (2008) *Risk Analysis: A Quantitative Guide*, 3rd ed. (John Wiley & Sons, Chichester, UK).

Weinschenk SM (2013) *How to Get People to Do Stuff: Master the Art and Science of Persuasion and Motivation* (Peachpit, San Francisco).

Wirth R (2000) CRISP-DM: Towards a standard process model for data mining. *Proc. Fourth Internat. Conf. Practical Appl. Knowledge Discovery Data Mining*, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133.