Chapter 5

Embracing the Data-Mining Process

In This Chapter

- Establishing a framework for data mining
- Drilling into the CRISP-DM process
- Establishing good habits

Data mining doesn't have official rules. You have tremendous flexibility to define and refine your own work methods. Still, you'll find benefits to understanding and following the approaches that work well for others.

The *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*) is the dominant process framework for data mining. It's an open standard; anyone may use it. This chapter explains each phase of the process.

Whose Standard Is It, Anyway?

The CRISP-DM process model is a step-by-step approach to data mining that was created by data miners for data miners. Participants from over 200 organizations (mainly a diverse group of businesses with an interest in using data mining internally or in promoting far-reaching use of data mining) provided input to develop the framework, which outlines key data-mining tasks in business terms and leaves users free to make their own choices about specific mathematical and computational approaches, and other technical matters.

TECHNICAL STUFF

The explanation of the CRISP-DM process in this chapter follows the original published version very closely. However, differences exist, such as changes in terminology or a diagram, intended to make the information clearer for data-mining novices. Also, the explanations in this book are briefer and lighter in style. If you would like to read the undiluted original (all 76 pages of it in small print), you can get it online (for free) at

```
ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/
                Documentation/14/UserManual/CRISP-DM.pdf
```

Approaching the process in phases

The CRISP-DM process model has six primary phases. These are

1. **Business understanding:** Get a clear understanding of the problem you're out to solve, how it impacts your organization, and your goals for addressing it.
2. **Data understanding:** Review the data that you have, document it, and identify data management and data quality issues.
3. **Data preparation:** Get your data ready to use for modeling.
4. **Modeling:** Use mathematical techniques to identify patterns within your data.
5. **Evaluation:** Review the patterns you have discovered and assess their potential for business use.
6. **Deployment:** Put your discoveries to work in everyday business.

Each of these phases involves several major tasks, and each task calls for several deliverables — primarily reports that summarize the work done and the information learned in that phase of the data-mining process. However, CRISP-DM does not define templates for these deliverables. You must plan and create them to suit the specific needs and style of your own workplace.

CRISP-DM defines the data-mining process primarily from a business standpoint. It tells you a lot about what you need to do, but it doesn't lay out all the technical details.

Cycling through phases and projects

Data mining is not something you do once and then forget. It's an ongoing cycle of activity. In any given project, you may address just a small element of a large and important problem, but you'll come back to that problem again and again with new projects. Because your work can also be applied to new projects, you'll revisit your past projects often, to see whether the models you developed in the past are still effective and to look for opportunities to improve on what you've done. Recycling your work in this way minimizes effort and helps you avoid confusion.

The CRISP-DM *process model* (not a mathematical model, but a set of guidelines for data-mining work) is a cycle often represented by a diagram like the one shown in Figure 5-1. Each project begins with business understanding and steps through each of the five phases of the process. Within the cycle, you find smaller cycles, so you may make several passes back and forth as you work to understand the business and the data, or to prepare data and build models. The cycle repeats as your project evaluation and experience during deployment add to your understanding of the business and inspire new projects.

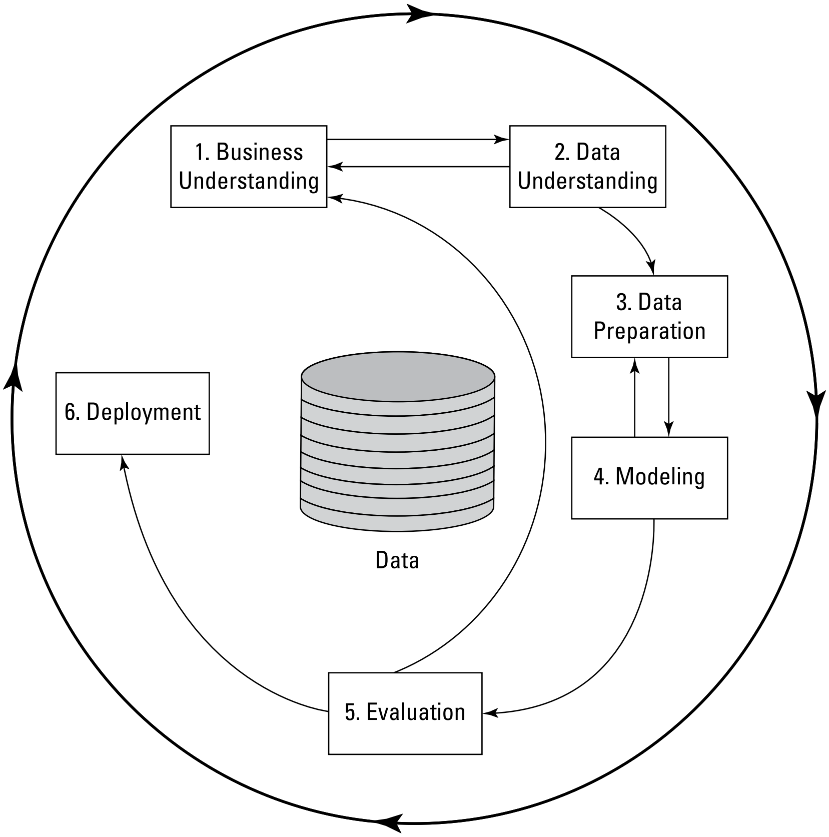

Figure 5-1: The CRISP-DM process model.

Documenting your work

When you are in the midst of a project, deeply involved with your data and the issues you've set out to address, you can easily get so familiar with the details that they seem obvious. You may not write things down because you don't see any need. Later, though, when you've moved on to other projects, spent time thinking about different datasets and different issues, and then returned to this project, you'll find that these details aren't obvious at all. You'll wonder what your sparse notes mean, be unsure of exactly where and how you got and prepared the data, and find other holes in your information.

Inadequate documentation leads to many problems. You may end up repeating work, and making others repeat work as well. You may fail to detect errors in your work. Your management or coworkers may become frustrated (and even angry) if you haven't prepared the documentation that they need. Failure to document your reasons for making certain decisions, or the proof that you met data privacy obligations, may even have legal consequences.

That's why much of the CRISP-DM process model is focused on reports and other documents that data miners create in the process of their work. These documents are your means of preserving information about what you've done so that you and others won't have to wonder later.

Business Understanding

In the first phase of a data-mining project, before you approach data or tools, you define what you're out to accomplish and define the reasons for wanting to achieve this goal.

The business understanding phase includes four *tasks* (primary activities, each of which may involve several smaller parts). These are

- Identifying your business goals
- Assessing your situation
- Defining your data-mining goals
- Producing your project plan

Task: Identifying your business goals

The first thing you must do in any project is to find out exactly what you're trying to accomplish! That's less obvious than it sounds. Many data miners have invested time on data analysis, only to find that their management wasn't particularly interested in the issue they were investigating. You must start with a clear understanding of

- A problem that your management wants to address
- The business goals

- Constraints (limitations on what you may do, the kinds of solutions that can be used, when the work must be completed, and so on)
- Impact (how the problem and possible solutions fit in with the business)

Deliverables for this task include three items (usually brief reports focusing on just the main points):

- **Background:** Explain the business situation that drives the project. This item, like many that follow, amounts only to a few paragraphs. Here's an example of a Background item:

  > *Our client, a regional planning commission, seeks to influence property use to enhance the quality of life for local residents. The planning commission has a broad charter which allows it to consider wide-ranging issues including employment, recreation, environment, and many other aspects of community life; however, the commission's role is purely advisory. It has a great deal of latitude to select issues for study, conduct research, and make policy recommendations to local lawmakers and staff, but does not have independent power to set regulations or influence property owners.*
  >
  > *Commission members (and others in local government and civic organizations) believe the best opportunity to influence property use occurs when the property changes hands. This implies that local government planning efforts can achieve greatest impact by focusing on properties which are about to change ownership. This poses a problem: the best time to act is before property changes hands, but the local government doesn't have dependable information about which properties are likely to be transferred. (Commercial real estate listings may be useful, but they do not cover all property transfers, and the best time to act may be before the property is listed.)*
  >
  > *Earlier research has identified a number of factors believed to indicate impending change of ownership; these include nonlocal ownership, multiple building code violations, and foreclosure, among others. While commissioners have good reason to believe that these factors influence the likelihood of property to change hands, their effects have not been quantified.*

- **Business goals:** Define what your organization intends to accomplish with the project. This is usually a broader goal than you, as a data miner, can accomplish independently. For example, the business goal might be to increase sales from a holiday ad campaign by 10 percent year over year.
- **Business success criteria:** Define how the results will be measured. Try to get clearly defined quantitative success criteria. If you must use subjective criteria (hint: terms like *gain insight* or *get a handle on* imply subjective criteria), at least get agreement on exactly who will judge whether or not those criteria have been fulfilled.

Task: Assessing your situation

This is where you get into more detail on the issues associated with your business goals. Now you will go deeper into fact-finding, building out a much fleshier explanation of the issues outlined in the business goals task.

Deliverables for this task include five in-depth reports:

- **Inventory of resources:** A list of all resources available for the project. These may include people (not just data miners, but also those with expert knowledge of the business problem, data managers, technical support, and others), data, hardware, and software.
- **Requirements, assumptions, and constraints:** Requirements will include a schedule for completion, legal and security obligations, and requirements for acceptable finished work. This is the point to verify that you'll have access to appropriate data!
- **Risks and contingencies:** Identify causes that could delay completion of the project, and prepare a contingency plan for each of them. For example, if an Internet outage in your office could pose a problem, perhaps your contingency could be to work at another office until the outage has ended.
- **Terminology:** Create a list of business terms and data-mining terms that are relevant to your project and write them down in a glossary with definitions (and perhaps examples), so that everyone involved in the project can have a common understanding of those terms.
- **Costs and benefits:** Prepare a cost-benefit analysis for the project. Try to state all costs and benefits in dollar (euro, pound, yen, and so on) terms. If the benefits don't significantly exceed the costs, stop and reconsider this analysis and your project.

TIP

Decision makers often feel more comfortable allotting resources to projects that reduce costs than those that aim to increase revenue, so always look for cost-savings potential, and state savings opportunities first in your costs and benefits report.

Task: Defining your data-mining goals

Reaching the business goal often requires action from many people, not just the data miner. So now, you must define your little part within the bigger picture. If the business goal is to reduce customer attrition, for example, your data-mining goals might be to identify attrition rates for several customer segments, and develop models to predict which customers are at greatest risk.

Deliverables for this task include two reports:

- **Data-mining goals:** Define data-mining deliverables, such as models, reports, presentations, and processed datasets.
- **Data-mining success criteria:** Define the data-mining technical criteria necessary to support the business success criteria. Try to define these in quantitative terms (such as model accuracy or predictive improvement compared to an existing method). If the criteria must be qualitative, identify the person who makes the assessment.

Task: Producing your project plan

Now you specify every step that you, the data miner, intend to take until the project is completed and the results are presented and reviewed.

Deliverables for this task include two reports:

- **Project plan:** Outline your step-by-step action plan for the project. Expand the outline with a schedule for completion of each step, required resources, inputs (such as data or a meeting with a subject matter expert), and outputs (such as cleaned data, a model, or a report) for each step, and dependencies (steps that can't begin until this step is completed). Explicitly state that certain steps must be repeated (for example, modeling and evaluation usually call for several back-and-forth repetitions).
- **Initial assessment of tools and techniques:** Identify the required capabilities for meeting your data-mining goals and assess the tools and resources that you have. If something is missing, you have to address that concern very early in the process.

Data Understanding

In the second phase of a data-mining project, conducted after you have defined goals and made a plan, you obtain data and verify that it is appropriate for your needs. You might identify issues that cause you to return to business understanding and revise your plan. You may even discover flaws in your business understanding, another reason to rethink goals and plans.

The data-understanding phase includes four *tasks*. These are

- Gathering data
- Describing data
- Exploring data
- Verifying data quality

Task: Gathering data

You've just set goals and defined a data-mining plan. Every step of the plan depends on having the right data. Better make sure that you really have that data!

Just one deliverable exists for this task: the initial data collection report. In your report, you need to verify that you have acquired the data or at least gained access to the data, tested the data access process, and verified that the data exists. You'll also need to load data into any tools that you will be using for data mining to verify that the tools are compatible with the data.

You may do a lot of work to assemble the data you need before you can write this report. First, you will make your plan, as follows:

- **Outline data requirements:** Create a list of the types of data necessary to address the data mining goals. Expand the list with details such as the required time range and data formats.
- **Verify data availability:** Confirm that the required data exists, and that you can use it. If some of the data you want is unavailable, decide how you will address that issue. Consider alternatives such as
    - Substituting with an alternative data source
    - Narrowing the scope of the project
    - Gathering new data
- **Define selection criteria:** Identify the specific data sources (databases, files, documents, and so on.) you will use. Within those sources, specify the tables, fields, and case ranges that are relevant to this project.

Once you've gone through these steps, you must actually obtain the data. At this stage, import the data into the data-mining platform you'll be using for the project to confirm that it is possible to do so and that you understand the process. In the course of this trial you may discover software (or hardware) limitations you had not anticipated, such as

- Limits on the number of cases or fields, or on the amount of memory you may use
- Inability to read the data formats of your sources
- Difficulty dealing with imperfections in the data (for example, you might encounter products that won't import or analyze incomplete datasets)

Finally, summarize the gathering process in a report. The report should describe your requirements, and explain in some detail exactly what data you have gathered and from what sources. Here you confirm that you have actually obtained the data and that it is compatible with your data-mining platform. If you have run into difficulties, you'll explain what they were and how you have addressed them (using alternative sources, revising plans, changing formats).

The deliverable for this task is just a simple report, but the work that you need to do before you can write that report won't be simple! Data access can be one of the most challenging and frustrating parts of the data-mining process, rife with both technical and business challenges.

Task: Describing data

Now that you have data, prepare a general description of what you have.

The deliverable for this task is the data description report. In it, you describe the source and formats of the data, the number of cases, the number and descriptions of the fields, and any other general information that may be important. You also make a brief evaluation of the suitability of the data for your data-mining goals. For example, verify that the data includes the fields that you expect and need to be there and sufficient cases for analysis.

Task: Exploring data

In this task, you examine the data more closely. For each variable, you look at the range of values and their distributions. You'll use simple data manipulation and basic statistical techniques for further checks into the data. Data exploration supports several purposes:

- Get familiar with the data.
- Spot signs of data quality problems.
- Set the stage for data preparation steps.

The deliverable for this task is the data exploration report. It's the place to document any hypotheses or initial findings that you have developed during data exploration. This report should include a more detailed description of the data than the data description report, including distributions, summaries, and any signs of data quality problems.

Task: Verifying data quality

You have the data and you've examined it, and now you have to determine whether it's good enough to support your goals. You will often have some quality problem to address yet still be able to move forward, but at times the data quality is so poor that it cannot support your plan and you'll have to look for alternatives. Some of the worst data problems would include

- The data you need doesn't exist. (Did it never exist, or was it discarded? Can this data be collected and saved for future use?)

- It exists, but you can't have it. (Can this restriction be overcome?)
- You find severe data quality issues (lots of missing or incorrect values that can't be corrected).

The deliverable for this task is the data quality report. This summarizes the data that you have, minor and major quality issues that you have found, and possible remedies for quality problems or alternatives (such as using an alternative data resource). If you are facing any really serious data quality issues and can't identify an adequate solution, you may have to recommend reconsidering goals or plans.

Data Preparation

Data miners spend most of their time on the third phase of the data-mining process: data preparation. Most data used for data mining was originally collected and preserved for other purposes and needs some refinement before it is ready to use for modeling.

The data preparation phase includes five *tasks*. These are

- Selecting data
- Cleaning data
- Constructing data
- Integrating data
- Formatting data

The CRISP-DM step-by-step guide does not explicitly mention datasets as deliverables for each of the data preparation tasks, but those datasets had darn well better exist and be properly archived and documented. Datasets won't correspond one-to-one with tasks, but information about the data used should be included in each deliverable report.

Task: Selecting data

Now you will decide which portion of the data that you have is actually going to be used for data mining.

The deliverable for this task is the rationale for inclusion and exclusion. In it, you'll explain what data will, and will not, be used for further data-mining work. You'll explain the reasons for including or excluding each part of the data that you have, based on relevance to your goals, data quality, and technical issues — such as limits to the number of fields or rows that your tools can handle, or the suitability of the data formats for your needs.

Task: Cleaning data

The data that you've chosen to use is unlikely to be perfectly clean (error-free). You'll make changes, perhaps tracking down sources to make specific data corrections, excluding some cases or individual cells (items of data), or replacing some items of data with default values or replacements selected by a more sophisticated modeling technique. You may choose to use only subsets of the data for all or some of your data-mining work.

The deliverable for this task is the data-cleaning report, which documents, in excruciating detail, every decision and action used to clean your data. This report should cover and refer to each data quality problem that was identified in the verify data quality task in the data-understanding phase of the process. You report should also address the potential impact on results of the choices you have made during data cleaning.

Task: Constructing data

You may need to derive some new fields (for example, use the delivery date and the date when a customer placed an order to calculate how long the customer waited to receive an order), aggregate data, or otherwise create a new form of data.

Deliverables for this task include two reports:

- **Derived attributes:** A report that describes what new fields (columns) you have constructed, how you did it, and why.
- **Generated records:** A report that describes what new cases (rows) you have constructed, how you did it, and why.

WARNING!

Although the merge data and format data tasks are listed last in this phase of the process, they don't always come last, and they may not come up just once. You might have to do some merging or reformatting early in the data preparation phase.

Task: Integrating data

Your data may now be in several disparate datasets. You'll need to merge some or all of those disparate datasets together to get ready for the modeling phase.

The deliverable for this task is the merged data. (And it would not hurt to document how the merge was performed.)

Task: Formatting data

Data often comes to you in formats other than the ones that are most convenient for modeling. (Format changes are usually driven by the design of your tools.) So convert those formats now.

The deliverable for this task is your reformatted data. (And a little report describing the changes you have made would be a smart thing to include.)

You should end the data preparation phase of the data-mining process with a dataset ready for modeling and a thorough report describing the dataset.

Modeling

This is the part of the process that most data miners like best. Your data is already in good shape, and now you can search for useful patterns in your data.

The modeling phase includes four tasks. These are

- Selecting modeling techniques
- Designing test(s)
- Building model(s)
- Assessing model(s)

Task: Selecting modeling techniques

The wonderful world of data mining offers oodles of modeling techniques, but not all of them will suit your needs. Narrow the list based on the kinds of variables involved, the selection of techniques available in your tools, and any business considerations that are important to you. (For example, many organizations favor methods with output that's easy to interpret, so decision trees or logistic regression might be acceptable, but neural networks would probably not be accepted.)

Deliverables for this task include two reports:

- **Modeling technique:** Specify the technique(s) that you will use.
- **Modeling assumptions:** Many modeling techniques are based on certain assumptions. For example, a model type may be intended for use with data that has a specific type of distribution. Document these assumptions in this report.

WARNING!

Statisticians are well-informed, strict, and fussy about assumptions. That's not necessarily true of data miners, and it's not a requirement to become a data miner. If you have deep statistical knowledge and understand the assumptions behind the models you select, you can be strict and fussy about assumptions. But many data miners, especially novice data miners, don't fuss much over assumptions. The alternative is testing — lots and lots of testing — of your models.

Task: Designing tests

The test in this task is the test that you'll use to determine how well your model works. It may be as simple as splitting your data into a group of cases for model training and another group for model testing. Training data is used to fit mathematical forms to the data model, and test data is used during the model-training process to avoid *overfitting:* making a model that's perfect for one dataset, but no other. You may also use *holdout data,* data that is not used during the model-training process, for an additional test.

The deliverable for this task is your test design. It need not be elaborate, but you should at least take care that your training and test data are similar and that you avoid introducing any bias into the data.

Task: Building model(s)

Modeling is what many people imagine to be the whole job of the data miner, but it's just one task of dozens! Nonetheless, modeling to address specific business goals is the heart of the data-mining profession.

Deliverables for this task include three items:

- **Parameter settings:** When building models, most tools give you the option of adjusting a variety of settings, and these settings have an impact on the structure of the final model. Document these settings in a report.
- **Model descriptions:** Describe your models. State the type of model (such as linear regression or neural network) and the variables used. Explain how the model is interpreted. Document any difficulties encountered in the modeling process.
- **Models:** This deliverable is the models themselves. Some model types can be easily defined with a simple equation; others are far too complex and must be transmitted in a more sophisticated format.

Task: Assessing model(s)

Now you will review the models that you've created, from a technical standpoint and also from a business standpoint (often with input from business experts on your project team).

Deliverables for this task include two reports:

- **Model assessment:** Summarizes the information developed in your model review. If you have created several models, you may rank them based on your assessment of their value for a specific application.
- **Revised parameter settings:** You may choose to fine-tune settings that were used to build the model and conduct another round of modeling and try to improve your results.

TIP

Data mining, like an onion, a Dobos torte, or a sedimentary rock, has lots of layers. When you are just getting started in data mining, you can start by leaving parameter settings at their default values (in fact, you might not even notice options unless you make an effort to look for them). As you get comfortable in your new data-mining career, it will make sense for you to find out about model parameters and know how you can use them. Your options will vary widely with the type of model and specific tool that you are using. The details are beyond the scope of this book, so when you are ready, refer to Chapter 19 for information about resources for discovering more about data mining.

Evaluation

You've explored data and you've found patterns, and now you have to ask: Are the results any good? You'll evaluate not just the models you create but also the process that you used to create them, and their potential for practical use.

The data-understanding phase includes three tasks. These are

- Evaluating results
- Reviewing the process
- Determining the next steps

Task: Evaluating results

At this stage, you'll assess the value of your models for meeting the business goals that started the data-mining process. You'll look for any reasons why the model would not be satisfactory for business use. If possible, you'll test the model in a practical application, to determine whether it works as well in the workplace as it did in your tests.

Deliverables for this task include two items:

- **Assessment of results (for business goals):** Summarize the results with respect to the business success criteria that you established in the business-understanding phase. Explicitly state whether you have reached the business goals defined at the start of the project.
- **Approved models:** These include any models that meet the business success criteria.

Task: Reviewing the process

Now that you have explored data and developed models, take time to review your process. This is an opportunity to spot issues that you might have overlooked and that might draw your attention to flaws in the work that you've done while you still have time to correct the problem before deployment. Also consider ways that you might improve your process for future projects.

The deliverable for this task is the review of process report. In it, you should outline your review process and findings and highlight any concerns that require immediate attention, such as steps that were overlooked or that should be revisited.

Task: Determining the next steps

The evaluation phase concludes with your recommendations for the next move. The model may be ready to deploy, or you may judge that it would be better to repeat some steps and try to improve it. Your findings may inspire new data-mining projects.

Deliverables for this task include two items:

- **List of possible actions:** Describe each alternative action, along with the strongest reasons for and against it.
- **Decision:** State the final decision on each possible action, along with the reasoning behind the decision.

Deployment

Deployment is where data mining pays off. It doesn't matter how brilliant your discoveries may be, or how perfectly your models fit the data, if you don't actually use those things to improve the way that you do business.

The deployment phase includes four tasks. These are

- Planning deployment (your methods for integrating data-mining discoveries into use)
- Planning monitoring and maintenance
- Reporting final results
- Reviewing final results

Task: Planning deployment

When your model is ready to use, you will need a strategy for putting it to work in your business.

The deliverable for this task is the deployment plan. This is a summary of your strategy for deployment, the steps required, and the instructions for carrying out those steps.

Task: Planning monitoring and maintenance

Data-mining work is a cycle, so expect to stay actively involved with your models as they are integrated into everyday use.

The deliverable for this task is the monitoring and maintenance plan. This is a summary of your strategy for ongoing review of the model's performance. You'll need to assure that it is being used properly on an ongoing basis, and that any decline in model performance will be detected.

Task: Reporting final results

Deliverables for this task include two items:

- **Final report:** The final report summarizes the entire project by assembling all the reports created up to this point, and adding an overview summarizing the entire project and its results.
- **Final presentation:** A summary of the final report is presented in a meeting with management. This is also an opportunity to address any open questions.

Task: Review project

Finally, the data-mining team meets to discuss what worked and what didn't, what would be good to do again, and what should be avoided!

This step, too, has a deliverable, although it is only for the use of the data-mining team, not the manager (or client). It's the experience documentation report. This is where you should outline any work methods that worked particularly well, so that they are documented to use again in the future, and any improvements that might be made to your process. It's also the place to document problems and bad experiences, with your recommendations for avoiding similar problems in the future.

REMEMBER

Data mining is a team activity. So if this process seems to include a lot of steps, realize that it may not be your personal responsibility to do every one of them, and that it's always appropriate to ask for help from others when you need it. (At the start of the project, you made a list of people who are resources for the data-mining project. That's your little directory of helpers!)