# https://www.dropbox.com/sh/3odimays52w21q2/AAAGTPGPTJO\_MLTg3k\_supAua?dl=0&preview=Data\_Science\_for\_Business\_v2.pdf

# https://www.scu.edu/ethics/internet-ethics-blog/should-twitter-suspend-donald-trumps-account/

# Chapter 1: Introduction to Data-Analytic Thinking

* Extensive investments in business infrastructure, improving data collection
* Every part of a businesss is open to data collection: operations, manufacturing, supply-chain, customer behaviour, marketing campaign performance, workflow procedures etc.
* Information now widely available on external events such as market trends, industry news and competitors movements.
* Ubiquitous and algoirthms have been developed that connect datasets to enabler deeper analysis.
* Datamining used for CRM to analyse behvvior – manage attrition, maximize customer value.
* Datamining is used for credit scoring and trading and in operations via. Fraud detection and workforce management.
* Datascience and Datamining are two separate things.
* Data science is a set of fundamental principles that guide the extraction of knowledge from data
* Data mining is the extraction of knowledge from data, via technologies that incorporate these principles. “Data science” is applied more broadly than data mining, but mining tech provide clearest illustrations of the principles of data science.

**Example: Hurricane:**Why data-drivne prediction might be useful in this scenario. -> predict who are in the path of the hurricane

Predict how many things Walmart will sell

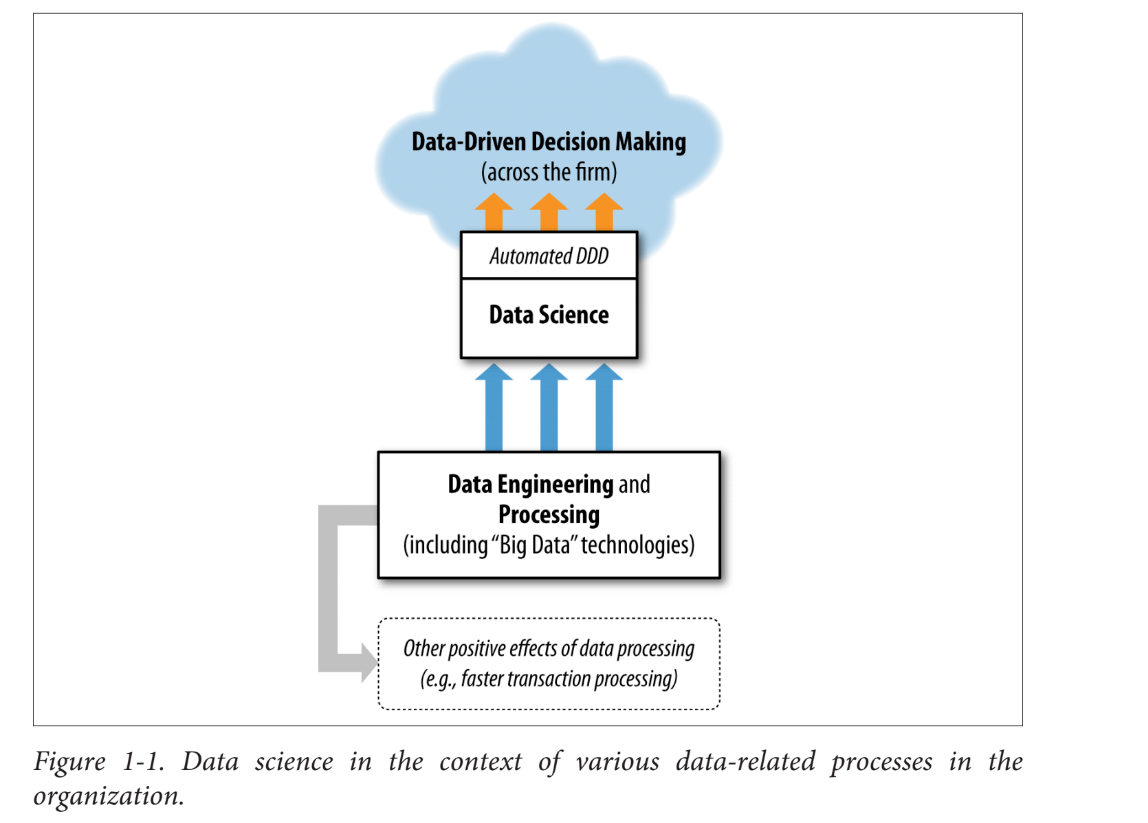
**Predicting Customer Churn:**\* Satured market

\*20% customer leaves after contract

\*what data will you ened etc.

**Data Science, Engineering and data-driven decision making.**

Data science involves principles, processes, techniques and understinging phenomena via automated analysis of data. The ultimate goal of data science as improving decision making, as this generally is of direct interest to a business.



Above figure illustrated data science in context of various other closely related and data-related processes in the organization.

It distinguishes data science from other aspects of data processing

**1. Data-driven decision-making**

Practice of basing decisions on the analysis of data, rather than on intuition.

Types: 1: Decisions for which discoveires need ot be made within data

Type 2: Decisions tha repeat, especially at massive scale and so decision-making can benefit from even small increases in decision-making accuracy based on data analysis.

**Data Processing and “Big Data”**

Data science needs access to data and benefits from sophisitacted data engeineering that data processing technologaies may facilitate.These technologies are not data science technologies per se.

“Big Data” Technologies (Such as Hadoop, HBASE and MongoDB)

Big data means datasets are too large for traditional data processing systems and require new processing technologies.

Big data technologies are used for: Data engineering, implementing data mining techniques, data processing in support of data mining.

**Big Data 1.0 to big data 2.0**

In Big Data 1.0, it is new and exciting. Businesses are trying to build capabilities to process large data.

Big Data 2.0 Golden era of data science, where processing data in a flexible fashion: “What can I now do that I couldn’t do before, or do better than I could do before?”

**Data and Data Science Capability as a Strategic Asset**

Fundamental principles of data science: Data, and the capability to extract useful knowledge from data, should be regarded as key strategic assets ( S 36) .

Businesses regard data analytics as realizing value from existing data, without considering if they have appropriate analytical talent.

**Does investing in Data Sceince pay off?**

**Amazon was able to gather data early on online customers, which created significant switching costs: cosnumers find value in the rankings and recommendations that Amazon provdes. They can therefore retain customers more easily and can even charge a premium. (S11)**

Data-Analytic Thinking

* **Data science is critical to understand, even if you don’t do it yourself due to strategic advantage.**
* **Businesses are increasingly driven by data analytics**
* **Managers need to be able to asses strategic dis-advantage, and therefore data. Data reaches into all business units**

**Fundamental concepts of data science:**

*Extracting useful knowledge from data to solve business problems*

*can be treated systematically by following a process with reasonably well-defined stages.*

* **Example: CRISP-DM is one codification of this process. This process provides a framework to structure our thinking about data analytics problems.**
* **For example, in actualy practice one sees analytical solutsions that are not based on careful analysis of the problem or are not carefully evaluated. Strucutred thinking about analytics emphasizes these often under-appreciated aspects of supporting decision-making with data. Human creativitry is necessary versus where high-powerted analytic tools can be brought to bear.**

*From a large mass of data, information technology can be used to*

*find informative descriptive attributes of entities of interest.*

* *In churn example: Customer is an entity of interest*
* *Each customer is described by attributes, such as usage, customer service etc.*
* *These attributes gives us info on customers likelihood of leaving the company.*
* *This process is reffered to as finding variables that “correlate” with curn.*
* *A business analyst may be able to hypothesize some and test them and there are tools to help facilitate this experimentation.*
* *An analyst can apply IT to automatically discover informative attributes, essentially oding large-scaled automated experimentation. Further, as we will see, this concept can be applied recursively to build models to predict churn based on multiple attributres.*

*If you look too hard at a set of data, you will find something—but*

*it might not generalize beyond the data you’re looking at*

This is called overfitting a dataset. Datamining techniques can be very powerful and the need to detect and avoid overfitting is one of the most important concepts to grasph when applying data mining to real problems. The concept of overfitting and its avoidance permeats data science processes, algorithms and evaluation methods.

*Formulating data mining solutions and evaluating the results*

*involves thinking carefully about the context in which they will be used.*

**We need to extract potentially useful knowledge. How do we formulate what is usefl? It depnds cirtically on the application in question. For our chunr-management example, how are we going to use the patterns extracted from historical data? Should the value of the customer be takne into account in addition to the likelihood of leaving? More generally, does the apttern lead to better decision than some reasonable alternative? How well would one have done by chance? How would would one do with a smart “Defualt” alternative?**

# Chapter 2: Business Problems and Data Science Solutions

**Fundamental concepts:**

*A set of canonical data mining tasks; The data mining process;*

*Supervised versus unsupervised data mining.*

* Important principle of data science, is that data mining is a processs with well-understood stages.
* Some involve IT with discovery of patterns from data, others require analysts creativity, business knowledge and common sense.
* Undertand the process to structure data mining project so they are systematic analysies rather than heroic endeavors.
* Data mining process breaks up the overall task of finding patterns from data, into a set of well-defined subtasks.
* Process as overaching framework -> Data mining process and common types of data mining tasks.

**From Business problems to data mining tasks**

* Each problem is unique – own combo of g oal,desires, constraints etc.
* There are sets of tasks that underlie business problems.
* Collab with stakeholders, data scientsis etc. decompose a problem into subtasks.
* Subtasks can be composed to solve overall problem
* Solution to any churn problem: Estimate from historical data, the probability of a customer terminating contract shortly after expiration.
* Probaility estimation fits the mold of one very common data mining task.
* Therefore: Crtical skill is ability to decompose data problems into pieces that each matches a known task for which tools are available. It avloid wasting resources and you can focus on important things. Human-creativity and intelligence must come into play, not just automatization.

1. **Classification and class probability estimation**

* Predict, for each data node, which class a data node belongs to. Clases are mutually exclusive. Example.” Among all customers, which are likely to respond?”. You have two clases: Will respond and will not respond (Mutually exclusive)
* Datamining procudes produces a model.
* Given a new indicidual, determines which class that individual belongs to.
* Closely related task si scoring or calss probability estimation.
* Scoring model applied to an individual produces instead of a class predicoitn a score representing the probability that hta tindividual bongs to each class.
* In a customer response scenario, a scoring model would be able to evaluate each individual customer and procude a score of how likely each is to respond to the offer. Classifcation and scoring are very closely related: asw shall see, a model that can do one usually can be modified to do the other.

1. **Regression (“Value estimation”)**

Attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.

En example regression question would be: “How much will a given customer use the service?”

* Prediction value: Service usage
* Model generated yb looking at similar indivudals in the population and historical usage.
* Regression procedure produces a model that estimates the value of the particiual variable specific to that individual.
* Regressoin is related to classificaition, but the two are different. Informally, clasiificatoin predicts whtether something will happen, regression predicts how much something will happen.

1. **Similarity matching:**

* Identify similar individuals based on data known about them.
* Can be used to find similar entties (ex: best customers based on firmographic data,
* One of the popular methods for making product recommendation (finding people hwho are similar to you in terms of the proucts they have liked or purchased)
* Similarity measures soltuions to other data minign tasks such as classification, regression and clustering.

1. **Clustering attempts to group individuals in a population together by their similarity.**

* Not driven by any specific purpose
* “Do our customers form antural groups or segments”
* Useful in preliminary domain exploration, to see which groups exist because these groups in turn may suggest other data mining tasks or approaches
* Clustering is used as input to decision-making processes. Focusong on questions such as “What products should we offer or develop?”

1. **Co-occurrence grouping (Also known as frequent itemset mining, association rule, discovery, market basket-analysis)**

* Attempts to find associations between entities based on transactions involving them
* “What items are commonly purchased together”
* Clusteirng looks at similarity between objects based on the attributes
* Co-occurrence grouping considers similarity of objects based on their appearing together in transactions
* Example: Recrods from IKEA may uncover tables are purchased with chairs, more frequently than expected.
* Recommendation systmes also perform affinity grouping, by finding pairs of chairs that are purchased **by the same kind of people. (X also bought Y)**

1. **Profiling (Known as behaviour description)**

* Attempt to characterize the typical behaviour of an individual, group or population.
* “What is the typical usage of this customer segment”
* Behavir may not have a simply description, profiling might require complex descritipn of night, weekend time etc, what they do when and how, how much.
* We can determine whatehter a new charge on the card fits the profile and use the dgree of mismatch as a supsion score and issue alarm if it is too high (Abuse of their card)

1. **Link prediction**

* Attempts to predict connectiosn between data items, by suggesting that a link should exist and estimating the strength of the link.
* Common in social networking systems (Since you share 10 friends, maybe add Karens’ fiend”
* Link prediction can estimate the strength of a link: Chance that a customer want to watch a movie that they have not watched. That we predixct should exist and should be strong

1. **Data reduction**

* Attempts to take a large set of data and replace it with a smaller set that contains the important information in the larger set.
* Smaller dataset is easier to process and may revela the information better.
* A massive dataset on consumer movie-viewing may be reducer to much smaller dataset revealing the tast eprefernces that are latent in the viewding data. (Viewer, genre preferences)
* Data reduction usualy involves loss of information, what is iomportant is the trade-off for improved insight.

1. **Casual modelling**

* Help us understand what events or actions influence others.
* Predicitve modelling to target advertisement to consumers – we observe that consumers purchase a higher rate subsequent to aving been targeted.
* Was it because of adverts or did the predictive models do a good job of identifiying the consumers wh would have purchased anayway.
* Techniques for causal modelling include those involving a sbustatial investment in data, such as randomized controlled exeriments (A/B tests) as wlel as methods for drawing causal conslusing from observational data.

Both experimental and observational methods for causal modeling generally can be viewed as “counterfactual” analysis: they attempt to understand what would be the difference between the situations—which cannot both happen —where the “treatment” event (e.g., showing an advertisement to a particular individual) were to happen, and were not to happen.

A careful data scientists should include with a causal conclusion, the exact assumptions that must be made in order for the causal conclusion to hold (There always are such assumptions – always ask)

* Book contains most fundamental data science principles – that together underline the types of tasks.
* Classification,regression, similarity matching, clustering.

**Supervervised versus unsupervised methods**

Questions you can ask about a customer population:

1. Do customers naturally fall into different groups? (is there a purpose or target specified?)

* If untargeted, then it is unsupervised methods you need to use

1. “Can we find groups of customers with high likelihoods of cancelling their service after contract expiration?”

* This is targeted, therefore we can segmentate for a reason – take action based onlikelihood of churn. This is supervised data mining problem.s
* Supervised vs unsupervised stems from Machine learning (Does the learning task involve examples or does the machine have to do it without any information about the purpose of the task)

Supervised = Specific target = supervised (Require different method than supervised), more useful results, specific purpose for grouping, predidcitng the target.

Unsupervised task = clustiering this produces groupin based on similarities, but no guarantee that the similarities are meaningful or will be useful for a purpose.

Conditions of supervised data minig:

* Specific target
* Must be data on the target (not enough that info exist in principle, it must exist in data)
* No go, if historical data is incomplete – target values cannot be provided.
* Acquiring data is a key data science investment, value for the target variable for an individual is called the individuals ‘labels’ – one must incur expense to actively label the data.

Otherwhise it is unsupervised data

Supervised methods with Classification, regression and causal modelling are solved with these methods

* Similarity matching, link prediction and data reduction

Unsupervised methods are generally

* Clustering, co-occurrence grouping and profiling.

**Main subclasses of supervised data mining:**

1. Classification
2. Regression

These two are distinguished by type of target. Regression involves numeric target, while classification involves a categorical target.

Supervised data minng questions that can be addressed:  
“ Will this customer prucahse Service A, if given incentive B?

* This is a binary target (The customer purchases or does not)

“Which service package (S1, S2, or none) will a customer likely purchase if given incentive I?”

This is also a classification problem, with a three-valued target.

“How much will this customer use the service?”

This is a regression problem because it has a numeric target. The target variable is the amount of usage (actual or predicted) per customer.

* For business problems we want numerical prediction over categorical target.
* IN churn example: yes/no prediction may not be sufficient.
* Want to model robability that the customer will continue (introducint numerical)
* Still considered classification modelling rather than regression, because underlyinig target is categorical. This is called “Class probability estimation”

Vital part in early stages of data mining process:

1. Decide if the line of attack will be supervised or unsupervised
2. If supervised, produce a precise definition of a target variable. This must be a specific quantity that will be the focus of the data mining. (Which we can obtain values for some example data)

**Distictions pertaining to mining data:**

**Difference between**

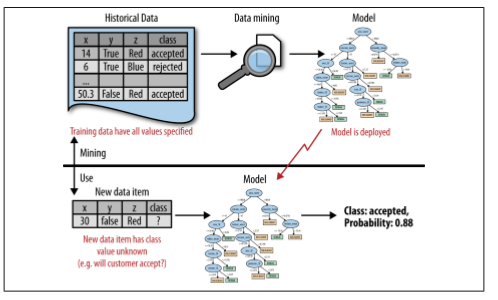
1. **Mining the data to find patterns and build models**
2. **Using the results of data minig.**

* **Students must NOT confuse these two when studying data science. Nor maangers when discussing businesss analytics. Use of data minig should influence and inform the data mining process, but the two are kept DISTINCT.**

**Example:**

**In our churn example, consider the deployment scenario in which the results will be used. We want to use the model to predict which of our customers will leave. Specifically, assume that data mining has created a class probability estimation model M. Given each**

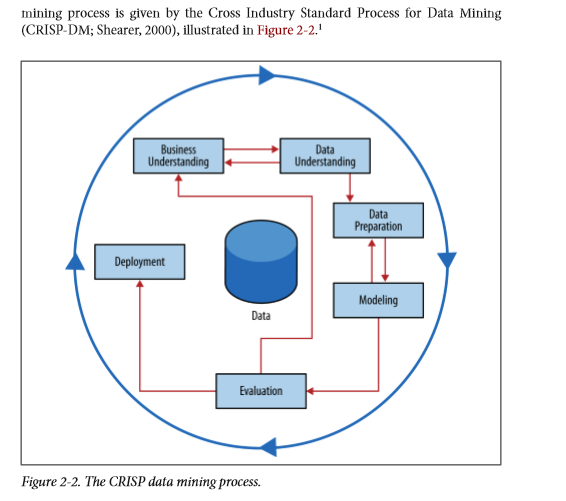
**Data Mining and Its Results | 25**



**Figure 2-1. Data mining versus the use of data mining results. The upper half of the figure illustrates the mining of historical data to produce a model. Importantly, the historical data have the target (“class”) value specified. The bottom half shows the result of the data mining in use, where the model is applied to new data for which we do not know the class value. The model predicts both the class value and the probability that the class variable will take on that value. existing customer, described using a set of characteristics, M takes these characteristics as input and produces a score or probability estimate of attrition. This is the use of the results of data mining. The data mining produces the model M from some other, often historical, data. Figure 2-1 illustrates these two phases. Data mining produces the probability estimation model, as shown in the top half of the figure. In the use phase (bottom half), the model is applied to a new, unseen case and it generates a probability estimate for it.**

## **The Data Mining Process ( CRISP)**

* Datamining is a craft.
* Involves application of science, technology and art.
* There is a well understodd process that places a structure on the problem.
* This allows consistence, repeatability and objectiveness. Useful codification of the data mining process is given by the cross indstury standard process for dataminng



* Interation is the rule!
* Going through once without solving the problems is not a failure
* Entire process is an exploration of the data and after the first iteration the team knows much more
* Next iteration can be much more well-informed

**Steps of CRISP for Datamining.**

1. **Business understanding:**

* Initially, vital to understand the problem to be solved
* Business projects seldom come pre-packaged as clear and umabigious data miing problems
* CRISP represents this as cycles within a cycle, it is not a lienar process. Initial formulation is incomplete or sub-optimal. Multiple iteration may be necessary for a good solution formlation to appear.
* Analysts creativity plays a large role
* High-level knowledge helps creative analysts see novel formulations
* Early stage: design a solution that takes advantage of these tools. (Structuring the prlbem such that one or more subproblems involve bulding models for classification, regression, probability estimation and so on.
* First stage: Design team should think carefully about the problem to be sovled and about the use scenario. (Chapter 7 & 11 )

**57**

1. **Data understanding**

* Comprise the available raw material that will be used for the solution
* Understand + and – of data because rearley is there an exact match with the problem
* Historical data are often collected for unrelated purposes to the business problem or for no purpose at all.
* Databases contain different info and may uncover different intersection populations with varying reliability.
* COST of data may vary, some will be available for free, others incur a cost to obtain. Some data may be purchased, other wont exist.
* A critical part of the data understanding phase is estimating the costs and benefits of each data source and deciding whether further investment is merited.

Even after all datasets are acquired, collating them may require additional effort. For example, customer records and product identifiers are notoriously variable and noisy

Cleaning and matching customer records to ensure only one record per customer is itself a complicated analytics problem

* As data understanding progresses, solution paths may change direction in response
* Team may fork, fraud detection provides an illustration of this.
* Data minig is used for fraud detection and many frad problems involve supervised data minig tasks.
* Consider task of catching credit card fraud: Charges show up on each customer account, so fraud charges are usually caught.
* We can assume that nearly all fraud is identified and reliably labelled, since the legitimate customer and the person perpetrating the fraud are different people with different goals.
* CC transactions have reliable labels (Fraud and legitimate) that serve as targets for a supervised technique.Poltiical, fake news? Are these labels maybe? Aswell as android, iphone. We catch the mall and see the origin. Are these good labels?

Who makes the tweets? We can’t know for sure.. do we use supervised or unsupervised techniques? If we do unsupervised, Do we make profiling of tweets origin with android / iphone? Do we cluster them and detect anomalities in the way the tweets are occurring? How about co-occurrence – maybe Trump always does something, after tweeting that others don’t?

* Insurance companies get fake claims – so we can’t have reliable labels!”!
* Scammers are subset of the legitimate users. No separate disinterested party will declare what the correct charges should be.
* Company has no reliable target variable, indicuatng fraud and a supervised learning approach that couldwork for credit card fraud is not applicable.
* This requires unsupervised approaches, suc has profiling, clustering, anomaly detection and co-occurrence grouping.

**In data understanding we need to dig beneath the surface to uncover the structure of the business problem and the data that are available**

match them to one or more data mining tasks for which we may have substantial science and technology to apply.

1. **Data preparation**

* Analytic technolgoies are powerful but impose requirements on the data.
* Data needs to be in a form different from how it is normally provided
* Conversion is necessary
* Data preparation phase proceeds along with data understanding – data are manipulated and converted into forms that yield better results (Think alteryx)

Examples:  
1. Convert data into tabular form

1. Remove missing values, convert data to different types
2. Data designed for symbolic / categorical data.
3. Data designed for numerical values

* Data scientists may spend time early to define variables used later in the process. This si the main point at which creativity, common sense and business knowledge come into play. Quality of data rests on how well the analyst structure the problems and craft the variables.

during data preparation is to beware of “leaks” (Kaufman et al. 2012). A leak is a situation where a variable collected in historical data gives information on the target variable—information that appears in historical data but is not actually available when the decision has to be made

example, when predicting whether at a particular point in time a website visitor would end her session or continue surfing to another page, the variable “total number of webpages visited in the session” is predictive. However, the total number of webpages visited in the session would not be known until after the session was over (Kohavi et al., 2000)—at which point one would know the value for the target variable!

Leakage must be considered carefully during data preparation, because data preparation typically is performed after the fact—from historical data.

1. Modelling

* Output of modelling is a model or pattern capturing regularites in the data
* Priamry stage where data mining techniques are pplied to the data
* Techniques and algorithsm is the part of the craft where most science and technology can be brought to bear.

1. Evaluation:

* Asses data mining esults – they need to be valid and reliable before moving on.
* Looking at datasets reveals patterns, but they may not survive careful scrutiny.
* We need confidence that the model and patternrs extracted from data are true regularites and not just idiosyncrasises or sample anomalies.
* It is possible to deploy results immediately after data mining, but is inadvisable. It is quicker,cheaper,easier and safer to test a model first in a laboratory stetting.
* Evaluation stage servers to help ensure that the model satisfies the business goals.
* Priamry goal of data science for business is to support decision making – we started the process by focusong on the business problem we would like to solve.
* a data mining solution is only a piece of the larger solution, and it needs to be evaluated as such
* A model may be extremely accurate (> 99%) by laboratory standards, but evaluation in the actual business context may reveal that it still produces too many false alarms to be economically feasible.

Evaluating the results of data mining includes both quantitative and qualitative assessments.

often stakeholders are looking to see whether the model is going to do more good than harm, and especially that the model is unlikely to make catastrophic

To facilitate such qualitative assessment, the data scientist must think about the comprehensibility of the model to stakeholders (not just to the data scientists). And if the model itself is not comprehensible (e.g., maybe the model is a very complex mathematical formula), how can the data scientists work to make the behavior of the model be comprehensible.

in some cases we may want to extend evaluation into the development environment, for example by instrumenting a live system to be able to conduct randomized experiments. In our churn example, if we have decided from laboratory tests that a data mined model will give us better churn reduction, we may want to move on to an “in vivo” evaluation, in which a live system randomly applies the model to some customers while keeping other customers as a control group

). We may also want to instrument deployed systems for evaluations to make sure that the world is not changing to the detriment of the model’s decision-making. For example, behavior can change—in some cases, like fraud or spam, in direct response to the deployment of models

**4 deployment**

the results of data mining and the data mining techniquesare put into use to realize some ROI

The clearest cases of deployment involve implementing a predictive model in some information system or business process.

a model for predicting the likelihood of churn could be integrated with the business process for churn management. for example, by sending special offers to customers who are predicted to be particularly at risk.

the data mining techniques themselves are deployed

example, for targeting online advertisements, systems are deployed that automatically build (and test) models in production when a new advertising campaign is presented

Two main reasons for deploying the data mining system itself rather than the models produced by a data mining system are

e(i) the world may change faster than the data science team can adapt, as with fraud and intrusion detection

(ii) a business has too many modeling tasks for their data science team to manually curate each model individual

it is not necessary to fail in deployment to start the cycle again. The Evaluation stage may reveal that results are not good enough to deploy, and we need to adjust the problem definition or get different data.

**Regression analysis:**

extracting patterns that will generalize to other data, and for the purpose of improving some business process. Typically, this will involve estimating or predicting values for cases that are not in the analyzed data

The topic of explanatory modeling versus predictive modeling can elicit deep-felt debate,5 which goes well beyond our focus. What is important is to realize that there is considerable overlap in the techniques used, but that the lessons learned from explanatory modeling do not all apply to predictive modeling.

**Machine Learning and data mining:**

* The collection of methods for extracting (predictive) models from data, now known as machine learning methods
* Machine Learning as a field of study arose as a subfield of Artificial Intelligence
* over the years this data analysis aspect of machine learning has come to play a very large role in the field.
* machine learning methods were deployed broadly, the scientific disciplines of Machine Learning, Applied Statistics, and Pattern Recognition developed close ties, and the separation between the fields has blurred.

**Techniques to answer business questions**

1. **Who are the most profitagble customer**
2. **Is there a difference between the profitable customers and the average ones**
3. **Who are these customers, can I characterize theM?**
4. **Will some new customers be profitable? How much revenue should I expect?**

The first, a classification question, may be phrased as a prediction of whether a given new customer will be profitable (yes/no or the probability thereof). The second may be phrased as a prediction of the value (numerical) that the customer will bring to the company. More on that as we proceed.

## Chapter 3: Introduction to predictive modelling: From correlation to supervised segmentation

delves into one of the main topics of data mining: predictive modeling

begin by thinking of predictive modeling as supervised segmentation

how can we segment the population into groups that differ from each other with respect to some quantity of interes

how can we segment the population with respect to something that we would like to predict or estimate

target of this prediction can be something we would like to avoid, such as which customers are likely to leave the company when their contracts expire

one of the fundamental ideas of data mining: finding or selecting important, informative variables or “attributes” of the entities described by the data

What exactly it means to be “informative” varies among applications, but generally information is a quantity that reduces uncertainty about something

Information reduces uncertainty (pirate example)

A key to supervised data mining is that we have some target quantity we would like to predict or to otherwise understand better.

Often this quantity is unknown or unknowable at the time we would like to make a business decision

such as whether a customer will churn soon after her contract expires, or which accounts have been defrauded

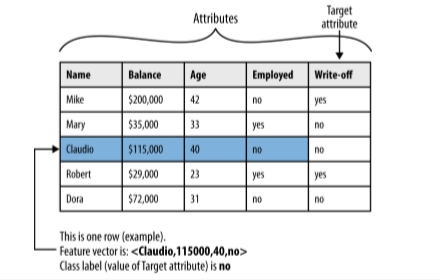
finding informative attributes: is there one or more other variables that reduces our uncertainty about the value of the target?

Finding informative attributes also is useful to help us deal with increasingly larger databases and data streams. Datasets that are too large pose computational problems for analytic techniques, especially when the analyst does not have access to highperformance computer

One tried-and-true method for analyzing very large datasets is first to select a subset of the data to analyze. Selecting informative attributes provides an “intelligent” method for selecting an informative subset of the data. In addition, attribute selection prior to data-driven modeling can increase the accuracy of the modelling

Tree induction incorporates the idea of supervised segmentation in an elegant manner, repeatedly selecting informative attributes

### Models, Induction and Prediction



A model is a simplified representation of reality created to serve a purpose. , a map is a model of the physical world

In data science, a predictive model is a formula for estimating the unknown value of interest: the target. The formula could be mathematical, or it could be a logical statement such as a rule. Often it is a hybrid of the two

Given our division of supervised data mining into classification and regression, we will consider classification models (and class-probability estimation models) and regression models.

in contrast to descriptive modeling, where the primary purpose of the model is not to estimate a value but instead to gain insight into the underlying phenomenon or process. A descriptive model of churn behavior would tell us what customers who churn typically look like.1

descriptive model must be judged in part on its intelligibility, and a less accurate model may be preferred if it is easier to understand

The difference between these model types is not as strict as this may imply; some of the same techniques can be used for both, and usually one model can serve both purposes (though sometimes poorly). Sometimes much of the value of a predictive model is in the understanding gained from looking at it rather than in the predictions it make

Supervised learning is model creation where the model describes a relationship between a set of selected variables (attributes or features) and a predefined variable called the target variable

model estimates the value of the target variable as a function (possibly a probabilistic function) of the features.

creation of models from data is known as model induction. Induction is a term from philosophy that refers to generalizing from specific cases to general rules (or laws, or truths).

models are general rules in a statistical sense (they usually do not hold 100% of the time; often not nearly), and the procedure that creates the model from the data is called the induction algorithm or learner

**Supervised segmentation**

intuitive way of thinking about extracting patterns from data in a supervised manner is to try to segment the population into subgroups that have different values for the target variable

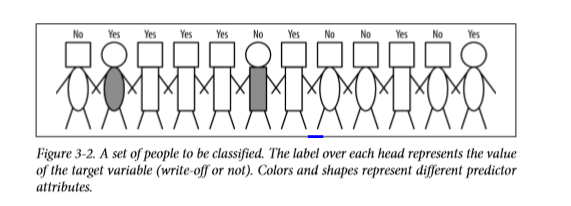
if the segmentation is done using values of variables that will be known when the target is not, then these segments can be used to predict the value of the target variable.

the segmentation may at the same time provide a human-understandable set of segmentation patterns

“middle-aged professionals who reside in New York City” is the definition of the segment

how can we judge whether a variable contains important information about the target variable? How much?

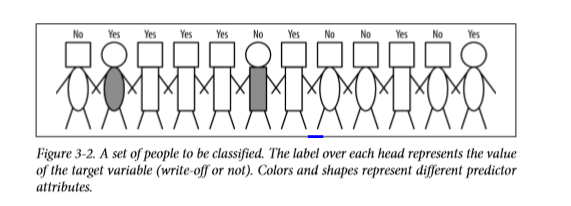
we might like to rank the variables by how good they are at predicting the value of the target.



one useful way to select informative variables

**Selecting Informative attributes**

**consider a binary (two class) classification problem, and think about what we would like to get out of it.**



**Figure 3-2 shows a simple segmentation problem: twelve people represented as stick figures. There are two types of heads: square and circular; and two types of bodies: rectangular and oval; and two of the people have gray bodies while the rest are white**

**• Attributes:**

**—head-shape: square, circular —body-shape: rectangular, oval —body-color: gray, white**

**• Target variable:**

**—write-off: Yes, No**

* Which way would be best to segment the people in groups? To distinguish write-poffs form non wirtes-offs.
* We want a group to be as pure as possible, homogeneous with the target variable. Either they are write-offs or not write offs. No inbetween.
* In real data we rarely expect to find a pure segment. We can reduce impurtity by example offer credit to tose with lower predicted write-off rates or offer different terms based on di
* fferent write-off rates.

Complications with impurity:1. Attributes rarely split a group perfectly. Even if one subgroup happens to be pure, the other may not. For example, in Figure 3-2, consider if the second person were not there. Then body-color=gray would create a pure segment (write-off=no). However, the other associated segment, body-color=white, still is not pure.

2. In the prior example, the condition body-color=gray only splits off one single data point into the pure subset. Is this better than another split that does not produce any pure subset, but reduces the impurity more broadly?

3. Not all attributes are binary; many attributes have three or more distinct values. We must take into account that one attribute can split into two groups while another might split into three groups, or seven. How do we compare these? 4. Some attributes take on numeric values (continuous or integer). Does it make sense to make a segment for every numeric value? (No.) How should we think about creating supervised segmentations using numeric attributes?

a formula that evaluates how well each attribute splits a set of examples into segments, with respect to a chosen target variable. Such a formula is based on a purity measure.

common splitting criterion is called information gain, and it is based on a purity measure called entropy

Entropy is a measure of disorder that can be applied to a set, such as one of our individual segments.

* Consider a set of properties of members of the set, each emmber has one of the properties.
* In supervised segmentation it will correspond to the values of the target variable
* Disorder corresponds to how impure the segment is with respect to the properties of interst
* A mixed up segment with lots of write-offs and lots of non-write etc would have high entropy

Equation 3-1. Entropy

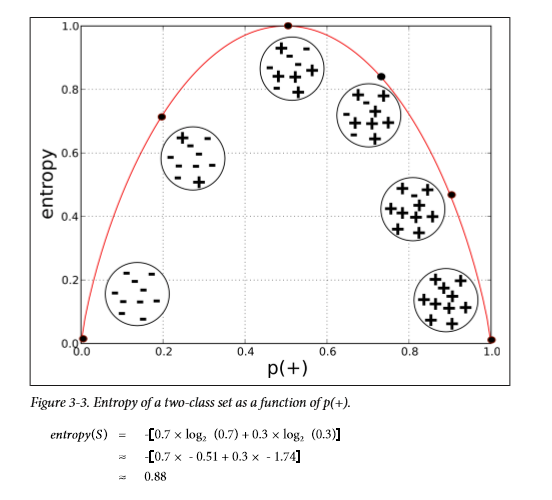
entropy = - p1 log(p1)- p2 log(p2)-

. Entropy is maximized at 1 when the instance classes are balanced

When all instances are positive, p+ = 1 and entropy is minimal again at zero

As a concrete example, consider a set S of 10 people with seven of the non-write-off class and three of the write-off class. So:

p(non-write-off)=7/10=0.7 p(write-off)=3/10=0.3



* Entropy is only part of the story
* Measure informative an attribute is with respect to our target.
* How much gain in information it gives us about the value of the target variable.
* Entropy only tells us how impure one individual subset is
* With entropy to measure disorder we can define information gain (IG) to measure how much an attribute improves(decreases) entropy over the segmentation it creates.
* Information gain measures the change in entropy due to any amount of new information being added