Data mining

Many definitions

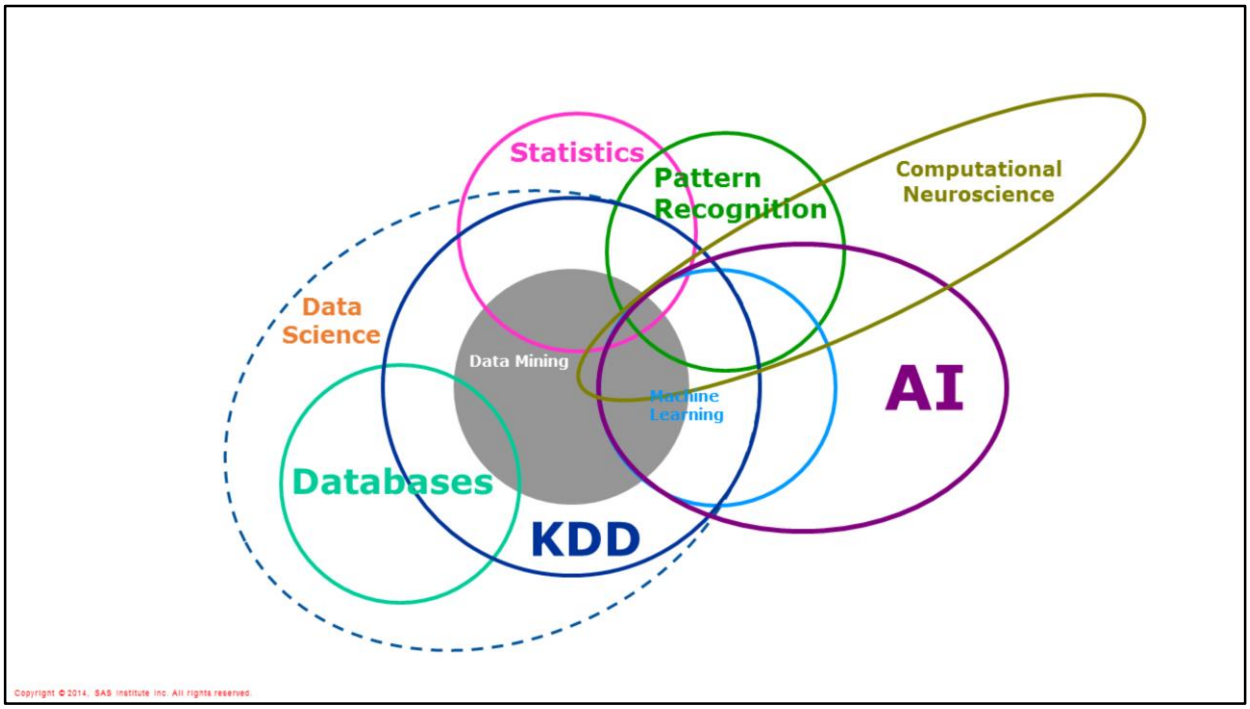
Introduction to Data Mining: "non-trivial extraction of implicit, previously unknown and potentially useful information from data."

Data mining differs from Statistics due to:

- It's focus on data storage and data manipulation methodologies
- It's focus on modeling methods that make few assumptions about the distribution of the training data, but that often have little theoretical support
- It's focus on commercial applications

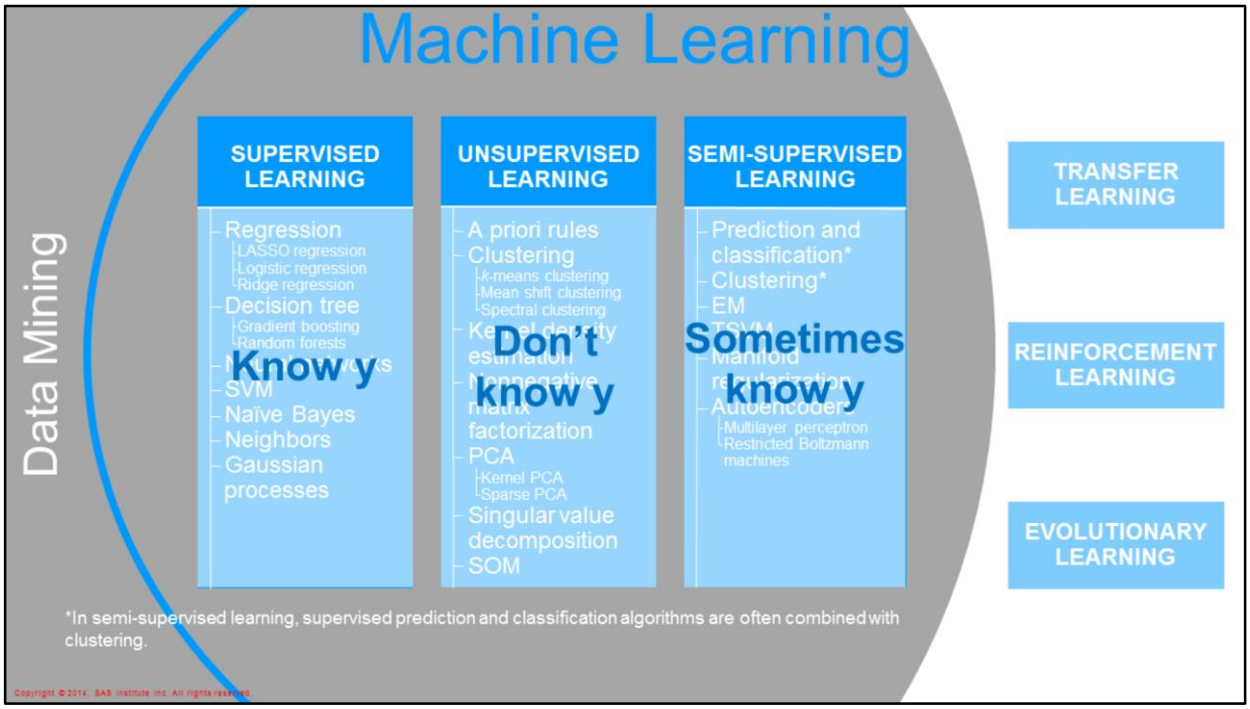
In a pop-culture sense, the terms "analytics", "big data", "data science", and "machine learning" are all basically synonyms of data analysis. "Data mining" was perhaps the precursor of these terms.

The data analysis field in general suffers from non-standard vocabulary issues. For instance, see the many different terms used for the rows and columns of a data set.



This busy graphic attempts to depict the multidisciplinary nature of many data analysis fields. If you're feeling like there is a lot overlap – there is! Statistical learning is a theoretical framework for machine learning; the statistical technique logistic regression is probably the most popular algorithm discussed in intro machine learning classes.

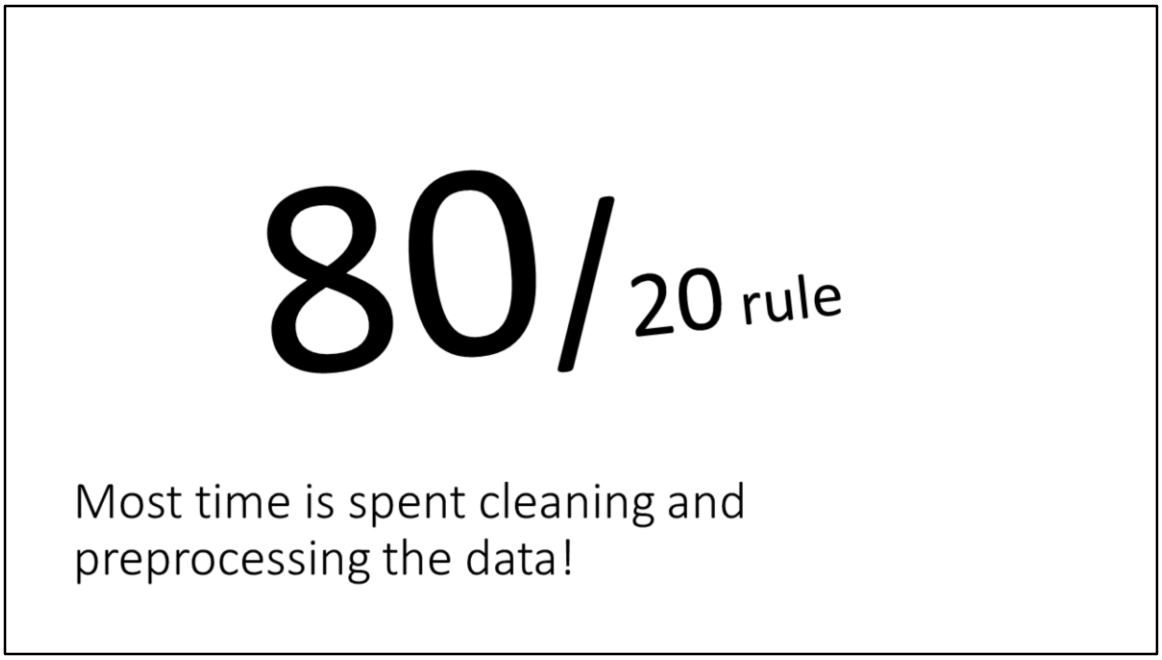
What many will recognize as machine learning, I would claim is within the intersection of data mining and machine learning, which also includes approaches from many other fields. It's a very rich area for data analysis algorithms. Let's take a closer look.



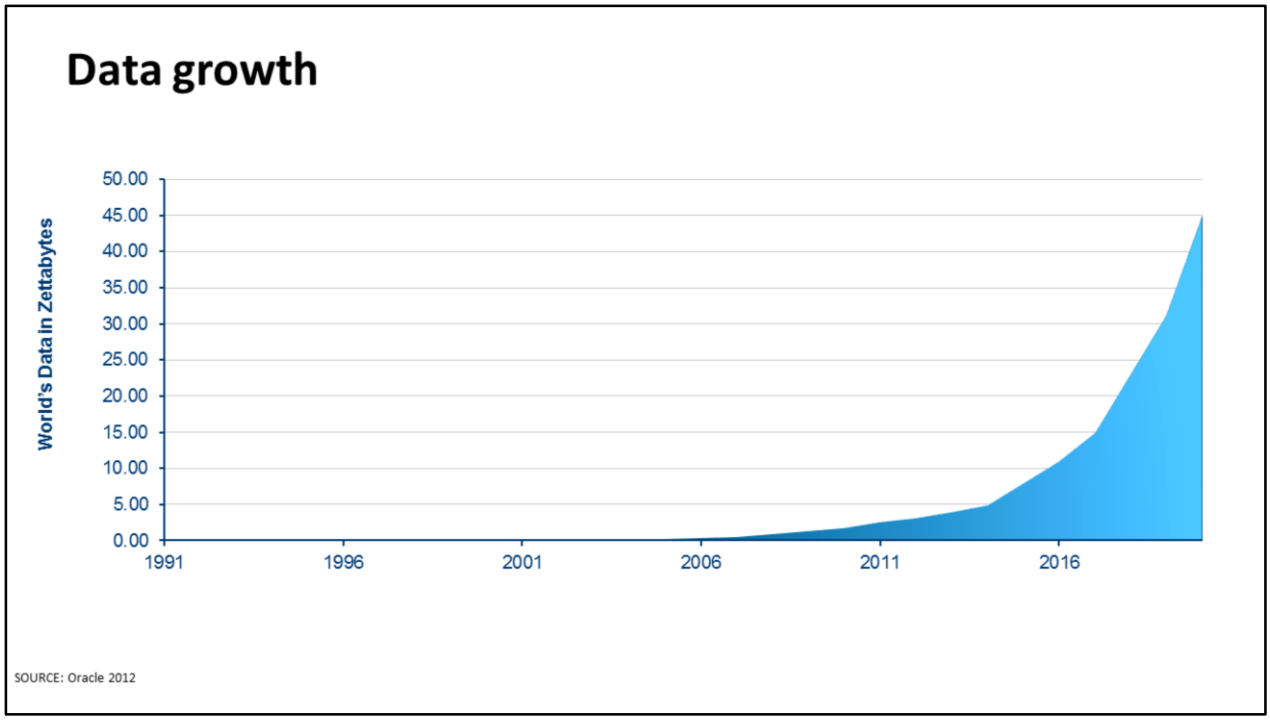
Supervised learning:
TARGET
Regression = simple supervised learning
Prediction – interval target
Classification – categorical target

Unsupervised learning:
NO TARGET
K-means = basic unsupervised learning
Clustering – groups similar rows of a data set
Feature selection – picking the most representative columns of a data set
Feature extraction – grouping the most representative columns in a data set into new columns

Semi-supervised learning:
Use some labeled data (hard to get) and a lot of unlabeled (easier to get) in combination with predictions/classification and clustering.
Some single algorithms, like autoencoders, are also directly capable of semi-supervised learning.



Most data mining education focuses on algorithms for modeling, while arguably the biggest part of the trade, cleaning and manipulating data sets, is often learned on the job.



Organizations have been accumulating massive amounts of data.

Structured data: columns and rows of numeric and well-formed character data
Transaction data: well-formed tuples, common in financial transactions: {UserID, ProductID, TimeStamp}

And importantly, diverse types of data:
Unstructured – images, text i.e. tweets, sound
Semi-structured – server logs, XML, JSON
Graph data – social networks, fraud or terrorism networks

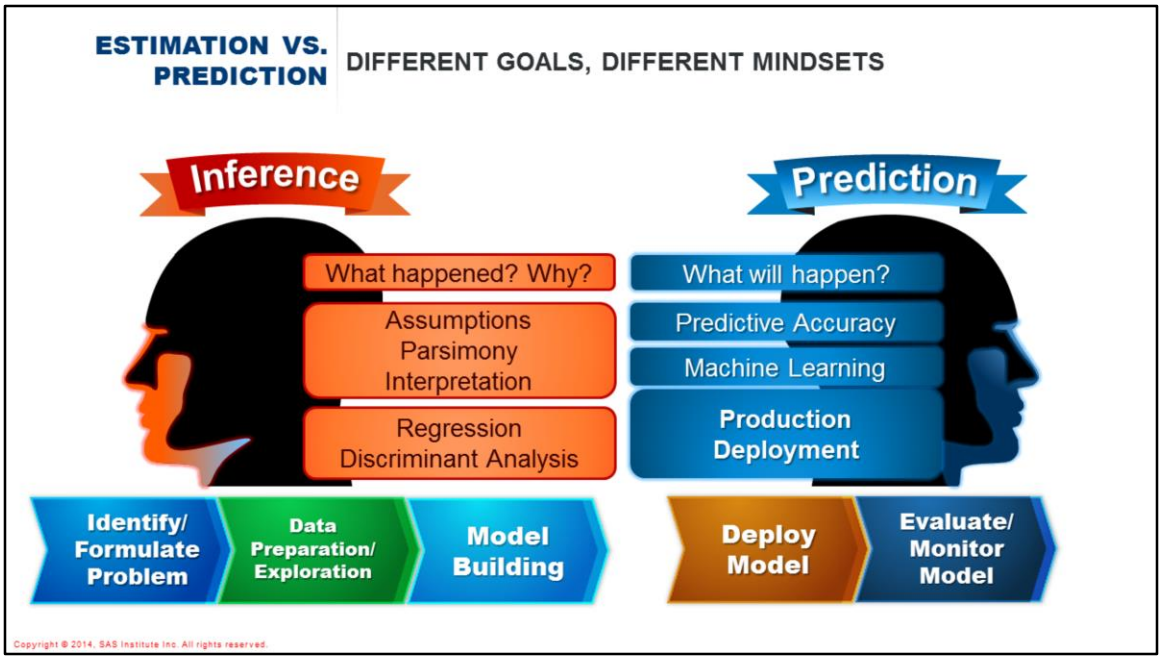
However, having a lot of data does not mean you have the right data! Criticisms of big data include:

1.) Sampling - it works; and in fact, done right, sampling and/or resampling can lead to better models than simply using all of your data.

2.) Redundancy - a lot the of data in a big data set can be redundant and therefore represented well by sampling weights or other more efficient mechanisms.

3.) Cost - the larger a data set is the more expensive it is to store: you need multiple servers, air conditioning, highly trained staff, etc. However, there is no guaranteed positive correlation between the size of a data set and the value of a data set. (Of course cloud storage is weakening this criticism ... but many organizations with data security concerns will continue to resist public cloud storage.)

4.) Design of experiment - a targeted, thoughtful experiment that collects the right information is much more efficient than just collecting all possible information.



The end goal of a data mining project is to create value, monetary or otherwise, from data.

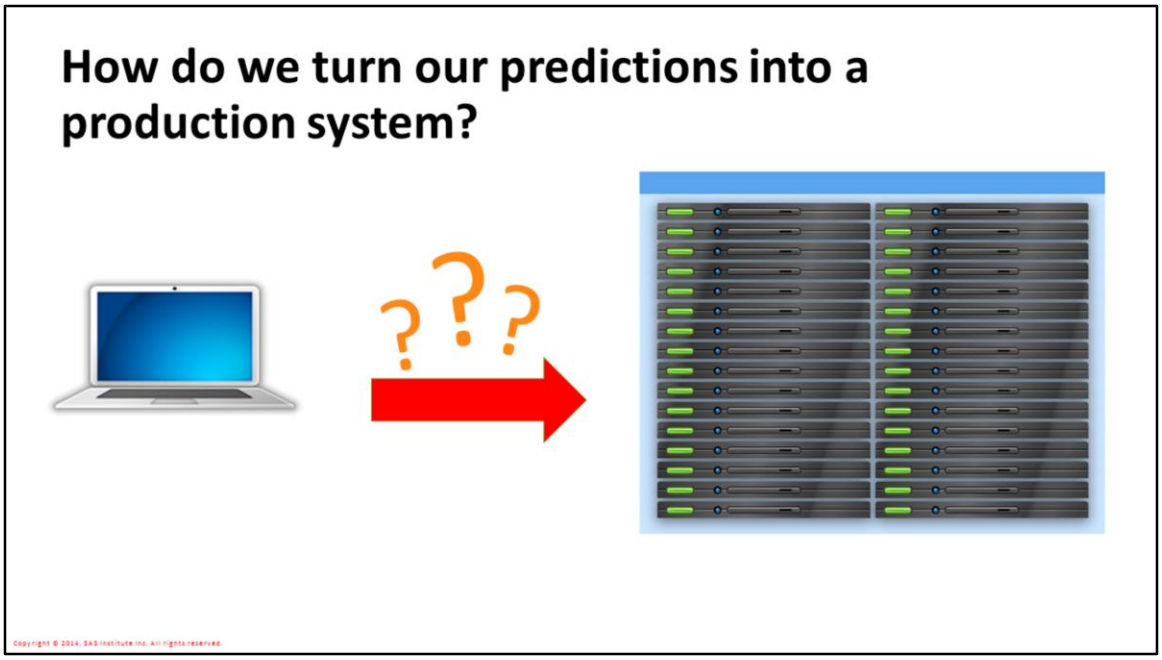
This can take the form of looking backward with data to understand the past, i.e. descriptive statistics and estimation of regression models, and statistical inference

OR

This can take the form of looking forward with data to make quantitative predictions, often using regression or more sophisticated machine learning models.

Both pursuits are legitimate and valuable, but they require different mindsets and different toolsets.

Another potential differentiator between traditional inferential approaches is that their output is meant to be interpretable by humans, and used for decision making by humans. Predictive machine learning methods were never meant to be interpretable by humans, but were meant to make the most accurate possible predictions on new data by using computers – so these techniques are basically meant for decision making by machines.



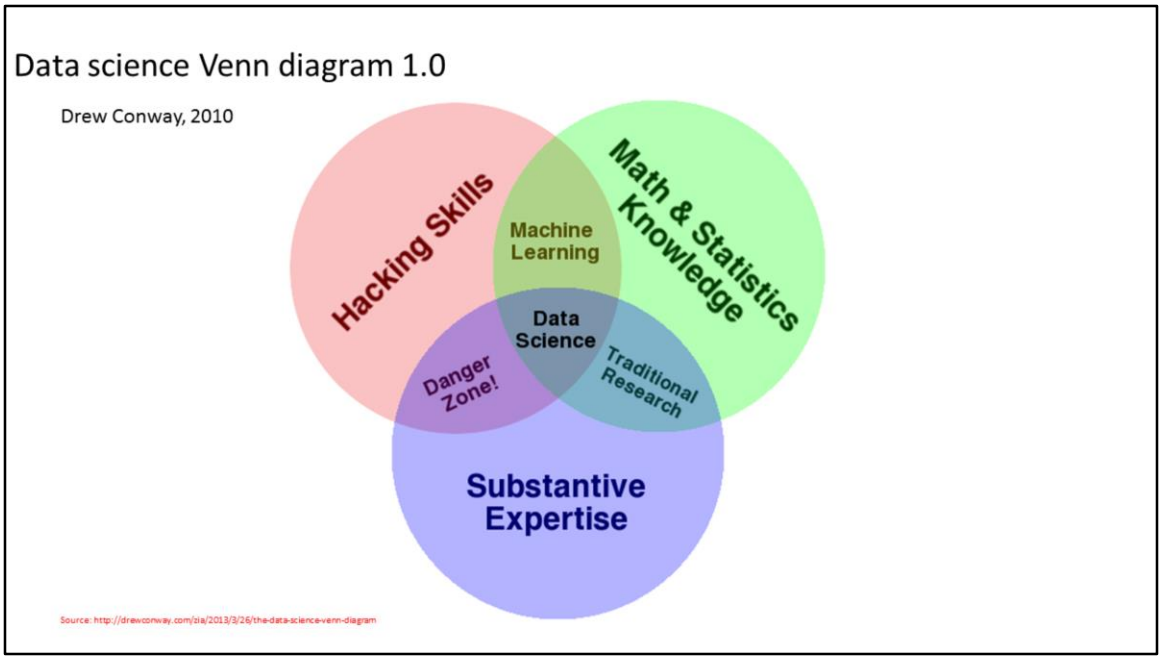
Making predictions on your laptop is only a good idea for a limited time in most cases.

If your model is really useful it needs to be used by your organization in an operational manner to make decisions quickly, if not automatically.

Moving the logic that defines all the necessary data preparation and mathematical expressions of a sophisticated predictive model from a development environment, like a personal laptop, into an operational computer system is one of the most difficult, tedious aspects of data mining. Mature successful, organizations are masters of this process – called “model deployment”, “deployment”, “model production” . While many immature organizations are often completely naïve about the process of model deployment.

To get some intuition for model deployment, think about a credit card company. They often use logistic regression models to automatically authorize each transaction. This logistic regression model probably starts out as Python, R or SAS code on an individual’s laptop or workstation. However, since this model must be used millions of times a day in a massive number of simultaneous authorization decisions that are guaranteed to be made in milliseconds, the model simply cannot be run on an individual’s laptop or workstation. Moreover, interpreted languages like Python, R, and SAS are probably too slow to guarantee millisecond response times – the model may need to be recoded into a compiled language, like C or maybe Java, wrapped in a web service, and/or containerized. This process of moving the model from an individual’s development environment, i.e. a laptop with R, to a large, powerful, secure database or server where it can be used by many mission-critical processes at once is “model deployment.”

Another example of model deployment is Gmail's spam detector. It is a predictive model that decides whether an incoming email message is spam or not. Think about the number of emails that must be screened. There is no one at Google running an R or Python script on their laptop for each piece of sent and/or received Gmail. The model just works automatically and must be part of some larger, powerful, and secure system that manages Gmail.



Because of all the different definitions and the massive breadth of the field of data analysis, I have come to prefer the blanket term “data science” … after several years of cynicism.

This picture is the best definition of a data science I know of.

Data scientists should use the scientific method. Otherwise they are just data alchemists, hoping to magically turn data into gold.