**Data Mining**Data Storage, organizationStatistical/Computational Problems, FrameworkBioinformaticsSupervised / Unsupervised LearningStatistical Learning - Predict / Identify / Estimate  
Machine Learning  
BigData  
looks at abstract data

**Unsupervised Learning**Learning from unstructured data

**Computer Vision  
pattern recognition**

**Supervised Learning**

**Knowledge discovery**  
How do you store, how do you use your data

**Learning from data**  
patterns, trends, associations  
Quantitative (stock prices) / Categorical (heart attack/no heart attack)  
Given a set of features predict quantitatively, categorically  
We have a training set of data, in which we observe the outcome and feature measure m

**Data Pipeline**Data collection – sensor, logging, experiments  
Encoding – cleaning categorizing, corrupt data, processing  
Processing – summarize it   
Exploration - Get to know your data – figure out numeric, categorical, types, risk  
Modeling – prediction, description, classification  
Communication – sharing with other people, commenting

Substantial Risk – compute probability of human extinction  
Data point are missing and choose to use -999 for those values  
Code Book – explains encode issue and dataset looks like  
If you build a model you get wrong prediction accuracy.

Unimodal – one bump

Bimodal -   
Symmetrical -  
Choosing a mean – symmetrical unimodal, you can with a bimodal.  
With bimodal you must do more analyzation  
  
Exploration – knowing how the data looks like with charts and graphs and see if your data is ok to not miss anything. Know why the oddities exist outliers can drive up the standard deviation.

Weight Blankets – metal chains inside blanket to see it improves sleeping.   
Huge improvement – had a bimodal distribution (two separate population)

Draw Historgram   
Regression Models  
Risk Research

Code Book

**Lectures**  
<http://www.math.csi.cuny.edu/~mvj/GC-DataMining/lectures/Lecture1.html#1>

**R Language  
Tutorial1 -** [**https://www.tutorialspoint.com/r/r\_csv\_files.htm**](https://www.tutorialspoint.com/r/r_csv_files.htm) **Tutorial2 -** [**http://www.r-tutor.com/elementary-statistics**](http://www.r-tutor.com/elementary-statistics) **Tutorial3 -** [**http://www.sthda.com/english/**](http://www.sthda.com/english/)

**The Elements of Statistical Learning**

**Chapter 1**

**Data Mining** includes applications in biology and medicine (bioinformatics). It is about extracting patterns, trends and learning from data.

**Supervised Learning** - involves predicting the outcome value based on input measure. The presence of the outcome variable to guid the learning process.

**Unsupervised Learning** - describes the association and patterns among input measures with not outcome measures. Observe on features and have no outcome measurement.

Indentifing Genes

Statistical Frameworks

Statistical Learning - learning from data

Applications - Medical (predicting heart attack), Financial (Predict Stocks), Digital Image (Identify Handwriting), Estimate

Quantitive Measurements - Mean, Max, Stock Price

Categorical Measurements - Yes / No,

Set of Features - diet

Training Set - observed data of the outcome and feature measurement for a set of objects through time.

Prediction Model (Learner) - is built using the Training Set

Classification Problem - a type of supervised problem that uses the outcome class (email / spam) if(%george < 0.6 & %you > 1.5) then spam else email. Scatter plots can show if the model is categorial.

Handwritten images (0-9)

Regression Problem - supervised learning problem with quantitative outcome measure  
Prostate Cancer

Unsupervised Learning Problem - Gene pro

Datasets used in the book: <http://www-stat.stanford.edu/ElemStatLearn>

**Engineering Statistics Handbook**

**http://www.itl.nist.gov/div898/handbook/eda/section4/eda4212.htm**

**EDA** – data analysis approach, Exploratory Data Analysis - allowing the data itself to reveal its underlying structure and model

**EDA Approach / Focus** – philosophy on a large venue of data analysis procedures

Statistical Graphics – collection of graphical techniques focusing on one data characterization.

**EDA** - a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret.

**EDA Techniques** - techniques are graphical in nature for reasons of open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data natural pattern-recognition capabilities.

plotting the raw data, plotting simple statistics, positioning such plots

**Data Analysis Approaches** – Classical, Exploratory (EDA), Bayesian. General science/engineering problem => science/engineering conclusions.

**Classical** - Problem => Data => Model => Analysis => Conclusions

the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model. Deterministic models include, for example, [regression models](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm) and [analysis of variance (ANOVA)](http://www.itl.nist.gov/div898/handbook/eda/section3/eda355.htm) models. The most common probabilistic model assumes that the errors about the deterministic **model** are normally distributed--this assumption affects the validity of the ANOVA F tests. the **focus** is on the model--estimating parameters of the model and generating predicted values from the model. **Techniques** - Classical techniques are generally [quantitative](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm) in nature. They include [ANOVA](http://www.itl.nist.gov/div898/handbook/prc/section4/prc42.htm), [t tests](http://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm), [chi-squared tests](http://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm), and [F tests](http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm). **Rigour** - the most important characteristic of classical techniques is that they are rigorous, formal, and "objective". **Data Treatment** - have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process. **Assumption** - The "good news" of the classical approach is that tests based on classical techniques are usually very sensitive--that is, if a true shift in location, say, has occurred, such tests frequently have the power to detect such a shift and to conclude that such a shift is "statistically significant". The "bad news" is that classical tests depend on underlying assumptions (e.g., normality), and hence the validity of the test conclusions becomes dependent on the validity of the underlying assumptions. Worse yet, the exact underlying assumptions may be unknown to the analyst, or if known, untested. Thus the validity of the scientific conclusions becomes intrinsically linked to the validity of the underlying assumptions. In practice, if such assumptions are unknown or untested, the validity of the scientific conclusions becomes suspect.

**EDA** - Problem => Data => Analysis => Model => Conclusions

the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what **model** would be appropriate. The Exploratory Data Analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data. For exploratory data analysis, the **focus** is on the data--its structure, outliers, and models suggested by the data. **Tehniques** - EDA techniques are generally [graphical](http://www.itl.nist.gov/div898/handbook/eda/section3/eda33.htm). They include [scatter plots, character plots](http://www.itl.nist.gov/div898/handbook/eda/section3/scatterp.htm), [box plots](http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm), [histograms](http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm), [bihistograms](http://www.itl.nist.gov/div898/handbook/eda/section3/bihistog.htm), [probability plots](http://www.itl.nist.gov/div898/handbook/eda/section3/probplot.htm), [residual plots](http://www.itl.nist.gov/div898/handbook/eda/section3/6plot.htm), and [mean plots](http://www.itl.nist.gov/div898/handbook/eda/section3/meanplot.htm). **Rigor** - very suggestive, indicative, and insightful about what the appropriate model should be. are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions. **Data treatment** - often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information. **Assumptions** - Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions.

**Bayesian** - Problem => Data => Model => Prior Distribution => Analysis => Conclusions

the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

**EDA vs Classical** - [Models](http://www.itl.nist.gov/div898/handbook/eda/section1/eda121.htm), [Focus](http://www.itl.nist.gov/div898/handbook/eda/section1/eda122.htm), [Techniques](http://www.itl.nist.gov/div898/handbook/eda/section1/eda123.htm), [Rigor](http://www.itl.nist.gov/div898/handbook/eda/section1/eda124.htm), [Data Treatment](http://www.itl.nist.gov/div898/handbook/eda/section1/eda125.htm), [Assumptions](http://www.itl.nist.gov/div898/handbook/eda/section1/eda126.htm)

**Models** - deterministic and probabilistic

**Deterministic Model** - [regression models](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm) and [analysis of variance (ANOVA)](http://www.itl.nist.gov/div898/handbook/eda/section3/eda355.htm) models

[**Regression Models**](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm) **–**

[**Analysis of Variance (ANOVA)**](http://www.itl.nist.gov/div898/handbook/eda/section3/eda355.htm) **-**

**Probabilistic Model** –

EDA vs Summary Analysis

**Summary Analysis -** is simply a numeric reduction of a historical data set which may then either replace the data set or be added to the data set in the form of a summary table.

EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data

EDM Goals

maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Insight implies detecting and uncovering underlying structure in the data To get a "feel" for the data, it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data. Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data. Graphics are irreplaceable--there are no quantitative analogues that will give the same insight as well-chosen graphics.

**EDM Graphics – provides** Statistics and data analysis procedures divided into [quantitative](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm)  and [graphical](http://www.itl.nist.gov/div898/handbook/eda/section3/eda33.htm)

|  |  |
| --- | --- |
| *Quantitative* | Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:   * [hypothesis testing](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm#TESTS) * [analysis of variance](http://www.itl.nist.gov/div898/handbook/prc/section4/prc42.htm) * [point estimates and confidence intervals](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm#INTERVAL) * [least squares regression](http://www.itl.nist.gov/div898/handbook/pmd/pmd.htm)   These and similar techniques are all valuable and are mainstream in terms of classical analysis. |
| *Graphical* | On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. These include:   * [scatter plots](http://www.itl.nist.gov/div898/handbook/eda/section3/scatterp.htm) * [histograms](http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm) * [probability plots](http://www.itl.nist.gov/div898/handbook/eda/section3/probplot.htm) * [residual plots](http://www.itl.nist.gov/div898/handbook/eda/section3/6plot.htm) * [box plots](http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm) * [block plots](http://www.itl.nist.gov/div898/handbook/eda/section3/blockplo.htm) |
| *EDA Approach Relies Heavily on Graphical Techniques* | The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of   * testing assumptions * model selection * model validation * estimator selection * relationship identification * factor effect determination * outlier detection   If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data. |

EDA Assumption

data from the process at hand "behave like":

1. random drawings;
2. from a fixed distribution;
3. with the distribution having fixed location; and
4. with the distribution having fixed variation.

univariate; that is, a single variable

response = deterministic component + random component

becomes

response = constant + error

|  |  |
| --- | --- |
| *Extrapolation to a Function of Many Variables* | The universal power and importance of the univariate model is that it can easily be extended to the more general case where the deterministic component is not just a constant, but is in fact a function of many variables, and the engineering objective is to [characterize and model the function](http://www.itl.nist.gov/div898/handbook/pmd/pmd.htm). |

|  |  |
| --- | --- |
| *Residuals Will Behave According to Univariate Assumptions* | The key point is that regardless of how many factors there are, and regardless of how complicated the function is, if the engineer succeeds in choosing a good model, then the differences (residuals) between the raw response data and the predicted values from the fitted model should themselves behave like a univariate process. Furthermore, the residuals from this univariate process fit will behave like:   * random drawings; * from a fixed distribution; * with fixed location (namely, 0 in this case); and * with fixed variation. |

|  |  |
| --- | --- |
| *Validation of Model* | Thus if the [residuals from the fitted model](http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd44.htm) do in fact behave like the ideal, then testing of underlying assumptions becomes a tool for the validation and quality of fit of the chosen model. On the other hand, if the residuals from the chosen fitted model violate one or more of the above univariate assumptions, then the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.    Predictability -probability statements not only about the process in the past, but also about the process in the future  If the four assumptions are not valid, then the process is drifting (with respect to location, variation, or distribution), unpredictable, and out of control. A simple characterization of such processes by a location estimate, a variation estimate, or a distribution "estimate" inevitably leads to engineering conclusions that are not valid, are not supportable (scientifically or legally), and which are not repeatable in the laboratory. |

|  |  |
| --- | --- |
| *EDA Questions* | Some common questions that exploratory data analysis is used to answer are:   1. What is a [typical value](http://www.itl.nist.gov/div898/handbook/eda/section3/eda351.htm)? 2. What is the [uncertainty for a typical value](http://www.itl.nist.gov/div898/handbook/eda/section3/eda356.htm)? 3. What is a [good distributional fit](http://www.itl.nist.gov/div898/handbook/eda/section3/ppccplot.htm) for a set of numbers? 4. What is a [percentile?](http://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm) 5. Does an [engineering modification have an effect?](http://www.itl.nist.gov/div898/handbook/eda/section3/bihistog.htm) 6. Does a [factor have an effect](http://www.itl.nist.gov/div898/handbook/eda/section3/blockplo.htm)? 7. What are the [most important factors?](http://www.itl.nist.gov/div898/handbook/eda/section3/dexmeanp.htm) 8. Are measurements coming from [different laboratories equivalent?](http://www.itl.nist.gov/div898/handbook/eda/section3/youdplot.htm) 9. [What is the best function for relating a response variable to a set of factor variables?](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm) 10. What are the [best settings for factors?](http://www.itl.nist.gov/div898/handbook/pri/section3/pri336.htm) 11. Can we separate [signal from noise in time dependent data](http://www.itl.nist.gov/div898/handbook/eda/section3/spectrum.htm)? 12. Can we extract any [structure from multivariate data](http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm)? 13. Does the data have [outliers](http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm)? |

Task 0 Tutorials

<https://stats.stackexchange.com/questions/157661/how-to-calculate-mean-median-mode-std-dev-from-distribution>

<http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r>

Group by Category

<https://www.r-bloggers.com/how-to-summarize-a-data-frame-by-groups-in-r/>

<https://www.tutorialspoint.com/r/r_csv_files.htm>

Linear Regression Model

<https://www.youtube.com/watch?v=owI7zxCqNY0>

<https://www.youtube.com/watch?v=ON_VvObtjz0>

https://www.tutorialspoint.com/r/r\_normal\_distribution.htm