Prof. Carlos Andrés Ferrero

Mineração de Dados

Introdução à Mineração de Dados

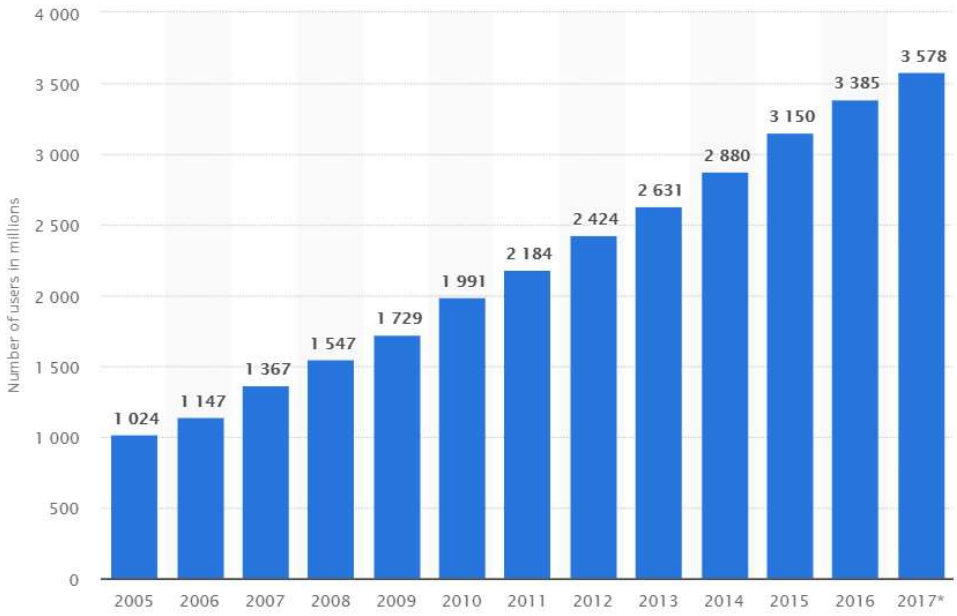
Conteúdo

- Processo de Descoberta de Conhecimento
- Aprendizado Supervisionado e Não-supervisionado
- Paradigmas de Aprendizado

Introdução

Número de Usuários na Internet

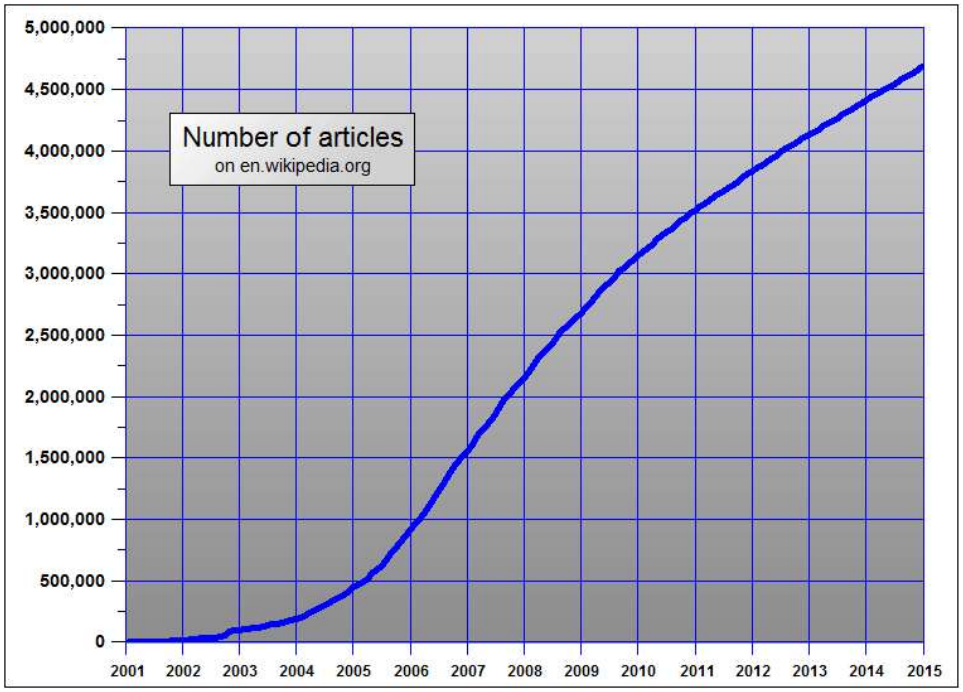
- A quantidade de usuários na Internet no mundo pasou de 15 milhões em 1995 para 2,8 bilhões em 2013-2014 (CASTRO e FERRERRI, 2016)
- Alguns dados dos últimos 12 anos mostram o crescimento



- Em julho de 2018 esse número foi para 4 bilhões (STATISTA, 2018)

Conteúdo no Wikipedia

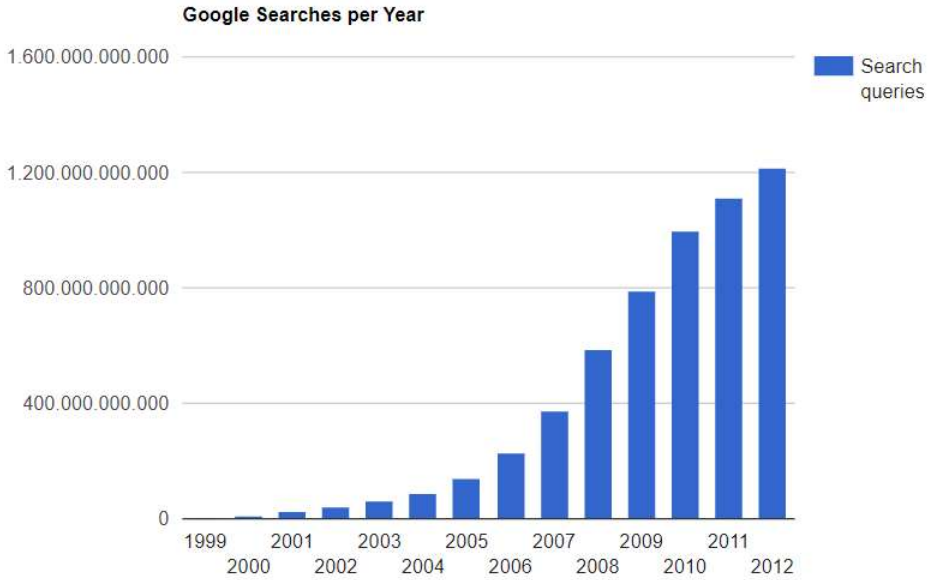
- A quantidade de conteúdo produzida no Wikipedia é enorme
- Em 2018 existem 5,5 milhões de artigos (apenas em Inglês)
- São produzidos cerca de 500 novos artígos por dia



By HenkvD (//commons.wikimedia.org/wiki/User:HenkvD) - Own work, CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0), Link (https://commons.wikimedia.org/w/index.php?curid=5702715)

Buscas no Google

- O Google processa cerca de 3,5 bilhões de buscas por dia (3,500,000,000)
- E cerca de 1,2 trilhões de buscas por ano (1,200,000,000,000)



Fonte: Internet Livestats (http://www.internetlivestats.com/google-search-statistics/#sources)

- Entre 15 a 20% dessas buscas nunca foram feitas antes.

Outras Estatísticas

- São escritos mais de 500 milhões de tuítes por dia
- São vistos mais de 200 milhões de horas de vídeo no Youtube por dia
- No Facebook são criado 5 perfís por segundo e existem cerca de 83 milhões de *fake profiles* (ZAFORIA.COM)
- ...

O que isto tem a ver com Mineração de Dados?

- A produção de dados pelos usuários de Internet é gigante
- O Google utiliza (e cria) técnicas de mineração de dados, como Recuperação de Informação (*Information Retrieval*) para recuperar resultados das nossas consultas. Isto é mais do que uma consulta SQL.
- Com os dados do Twitter é possível ver quais são os *hot topics* nas diferentes regiões do mundo e inferir diferentes sentimentos dos usuários (ou grupos de usuários) usando técnicas de mineração de textos (*text mining*), chamada de análise sentimento (*sentiment analysis*)
- O Facebook usa mineração de dados para aprender sobre *fake profiles* e poder dar um grau de probabilidade de que um perfil de facebook trata-se de um *fake*. A classificação é um exemplo de técnica utilizada para esse fim.

Áreas de Conhecimento envolvidas

- Visualização
- Estatística
- Matemática
- Engenharia
- Inteligência Artificial
- Banco de Dados
- Sistemas de Informação
- ...

- Do ponto de vista de um *data miner* em geral não precisa conhecer todas essas áreas do conhecimento profundamente antes de começar a prática de minerar dados, mas precisamos ter uma ideia dessas disciplinas para poder aprender com a experiência.
- Por exemplo, utilizar média e desvio padrão, dois temas fundamentais da Estatística, é vital para analisar dados, para entender a precisão de um *preditor*, entre outros.
- Mais atualmente, estamos escutando os termos de *Data Science* e *Data Scientist*, os quais demonstram claramente a importancia da Análise de Dados de uma forma geral no contexto de *Big Data*.
- Esses termos podem ser entendidos como uma evolução de data mining, onde procura-se analisar dados em diferentes formatos (convencionais, texto, imagens, temporais, entre outros) e, preferencialmente, de maneira conjunta, para extrair novas informações e conhecimentos, que sejam úteis para a sociedade.

Dados, Informação e Conhecimento

Uma visão prática (CASTRO e FERRARI, 2016):

- Dados
  - 1000 milibares
  - 5,1 m/s; 95 graus
  - 30 graus centígrados
  - poucas
  - 1000 metros

- Informação
  - Pressão atmosférica = 1000 milibares
  - Velocidade e direção do vento = 5,1 m/s; 95 graus
  - Temperatura do ar = 30 graus centígrados
  - Nuvens = poucas
  - Visibilidade = 1000 mts

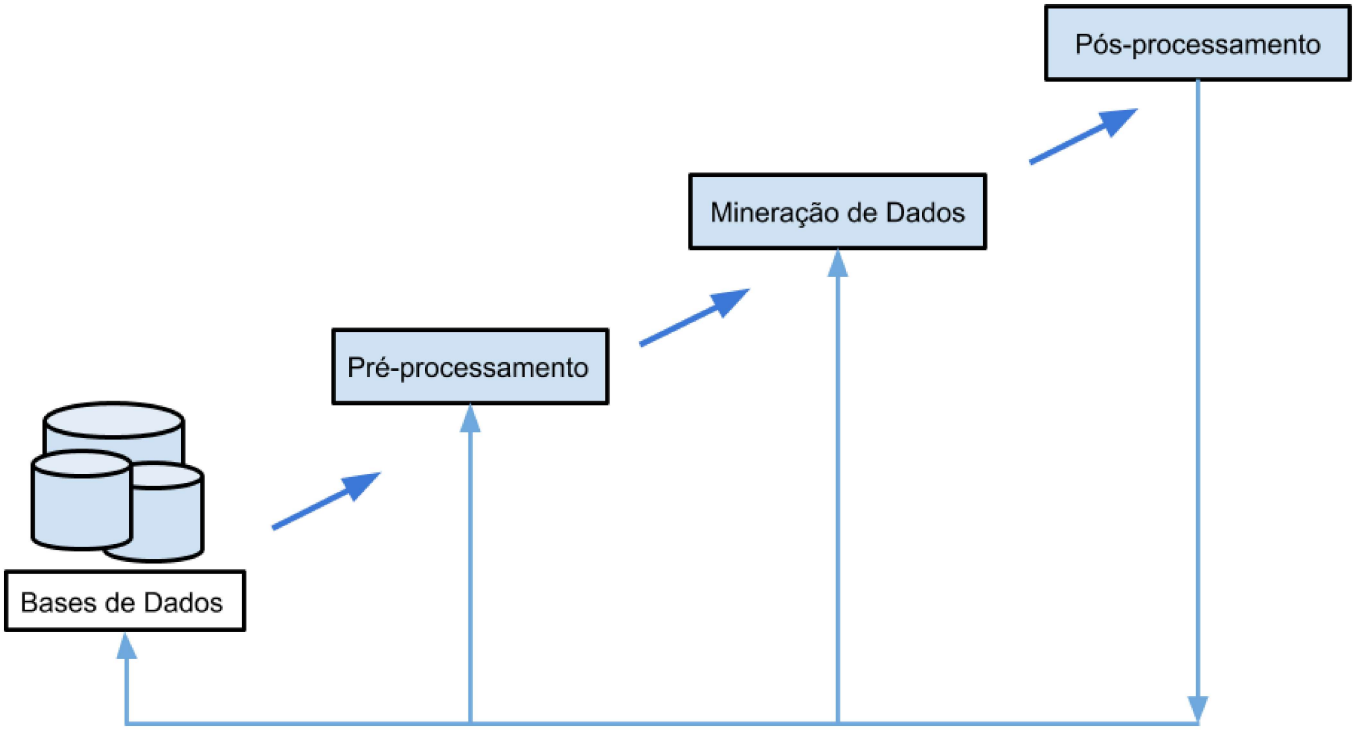
- Conhecimento
  - A probabilidade de chuva é baixa, portanto, posso ir à praia

- Definições:
  - **Dado**: símbolos e signos não estruturados, como os valores em tabelas e documentos
  - **Informação**: consistem nos dados interpretados, agregando significado e utilidade aos dados
  - **Conhecimento**: é algo que permite auxiliar em processos de tomada de decisão

Processo de Descoberta de Conhecimento em Bases de Dados

- Também chamado de KDD (*Knowledge Discovery in Databases*), foi proposto por Piatetski-Shapiro em 1989
- Esse processo tem como principal objetivo extrair padrões intrínsecos nos dados e apresentá-los de forma a facilitar a sua assimilação como conhecimento, bem como da utilização para outros fins

Etapas do Processo de Descoberta de Conhecimento (KDD)



Pré-processamento

- Visa preparar os dados para uma análise eficiente e eficaz
- Algumas tarefas de pré-processamento:
  - Limpeza (remoção de ruídos e dados inconsistentes)
  - Integração (combinação de dados obtidos a partir de múltiplas fontes)
  - Seleção de Dados ou Redução (escolha dos dados relevantes para análise)
  - Transformação (transformação dos dados para um formato adequado para análise)

Mineração de Dados ou Extração de Padrões

- Aplicação de algoritmos capazes de extrair padrões e/ou informações e conhecimento relevantes a partir dos dados pré-processados
- Algumas tarefas de mineração:
  - Análise descritiva estatística
  - Agrupamento
  - Modelos de Predição (classificação e regressão)
  - Associação (extração de regras de associação)
  - Detecção de Anomalias

Pós-procesamento ou avaliação de modelos

- Avaliação da qualidade dos modelos ou conhecimento extraído na etapa de Mineração de Dados
- Algumas tarefas de pós-processamento:
  - Verificar a precisão dos modelos construídos em novos dados
  - Identificar conhecimentos verdadeiros e conhecimentos triviais
  - Identificar problemas no pré-processamento ou mineração que precisem ser modificados para a melhoría do processo de descoberta de conhecimento

Um exemplo prático

- Estudo de um conjunto de dados de flores (*iris*)
- O dataset está constutuído de 150 flores (50 flores de cada espécie). Cada flor está representada por 4 atributos numéricos:
  - Sepal-length
  - Sepal-width
  - Petal-length
  - Petal-width

- As espécies ou *classes* são: *iris-versicolor*, *iris-virginica*, *iris-setosa*. Um exemplo dessas flores está nesta imagem:

Fonte: http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/ (http://suruchifialoke.com/2016-10-13-machine-learning-tutorial-iris-classification/)

Uma visão geral do conjunto de dados coletado

In [3]:
```
library(data.table)
dt.iris <- data.table(iris)
dt.iris.text <- dt.iris[,.(Exemplos=paste0("Sepal.Length ",Sepal.Length,
                                           ", Sepal.Width ",Sepal.Width,
                                           ", Petal.Length ",Petal.Length,
                                           ", Petal.Width ",Petal.Width,
                                           ", and Species ",Species) )]
```

In [4]:
```
head(dt.iris.text)
```

Exemplos
Sepal.Length 5.1, Sepal.Width 3.5, Petal.Length 1.4, Petal.Width 0.2, and Species setosa
Sepal.Length 4.9, Sepal.Width 3, Petal.Length 1.4, Petal.Width 0.2, and Species setosa
Sepal.Length 4.7, Sepal.Width 3.2, Petal.Length 1.3, Petal.Width 0.2, and Species setosa
Sepal.Length 4.6, Sepal.Width 3.1, Petal.Length 1.5, Petal.Width 0.2, and Species setosa
Sepal.Length 5, Sepal.Width 3.6, Petal.Length 1.4, Petal.Width 0.2, and Species setosa
Sepal.Length 5.4, Sepal.Width 3.9, Petal.Length 1.7, Petal.Width 0.4, and Species setosa

Pré-processamento

Extração dos dados do texto para um formato atributo-valor. Esse formato é representado por uma tabela onde as *linhas* são as instâncias (ou exemplos) de flores e as *colunas* os atributos das flores. Definimos os primeiros 4 atributos como numéricos e o atributo *Species* como categórico.

In [5]:
```
head(dt.iris, 10)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

Descrição de dados

In [10]:
```
dt.iris.melt <- melt(dt.iris, id.vars="Species")
```

In [11]:
```
dt.iris.melt[,.(mean = mean(value), sd = sd(value)), .(variable) ]
```

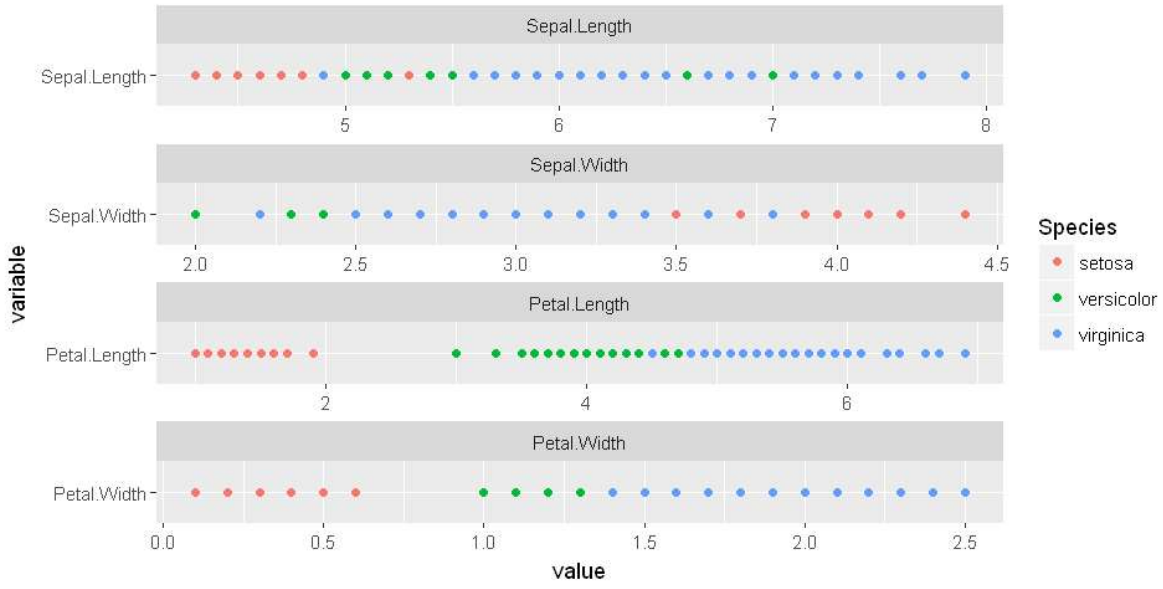
variable	mean	sd
Sepal.Length	5.843333	0.8280661
Sepal.Width	3.057333	0.4358663
Petal.Length	3.758000	1.7652982
Petal.Width	1.199333	0.7622377

Visualização de dados

- *Análise da distribuição dos valores dos atributo por classe*

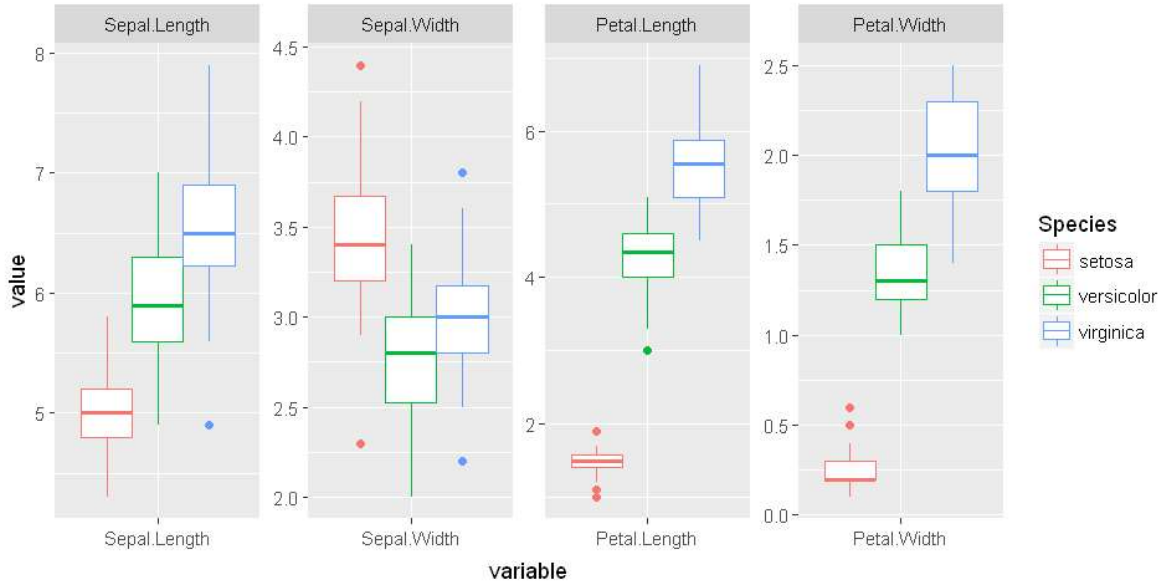
In [14]:
```
library(ggplot2)
options(repr.plot.width=8, repr.plot.height=4)
```

In [15]:
```
ggplot(data=dt.iris.melt, aes(y=variable, x=value, color=Species)) + geom_point() + facet
_wrap(~variable, ncol=1, scales = "free")
```



- *Diagrama de Caixas dos atributos por classe*

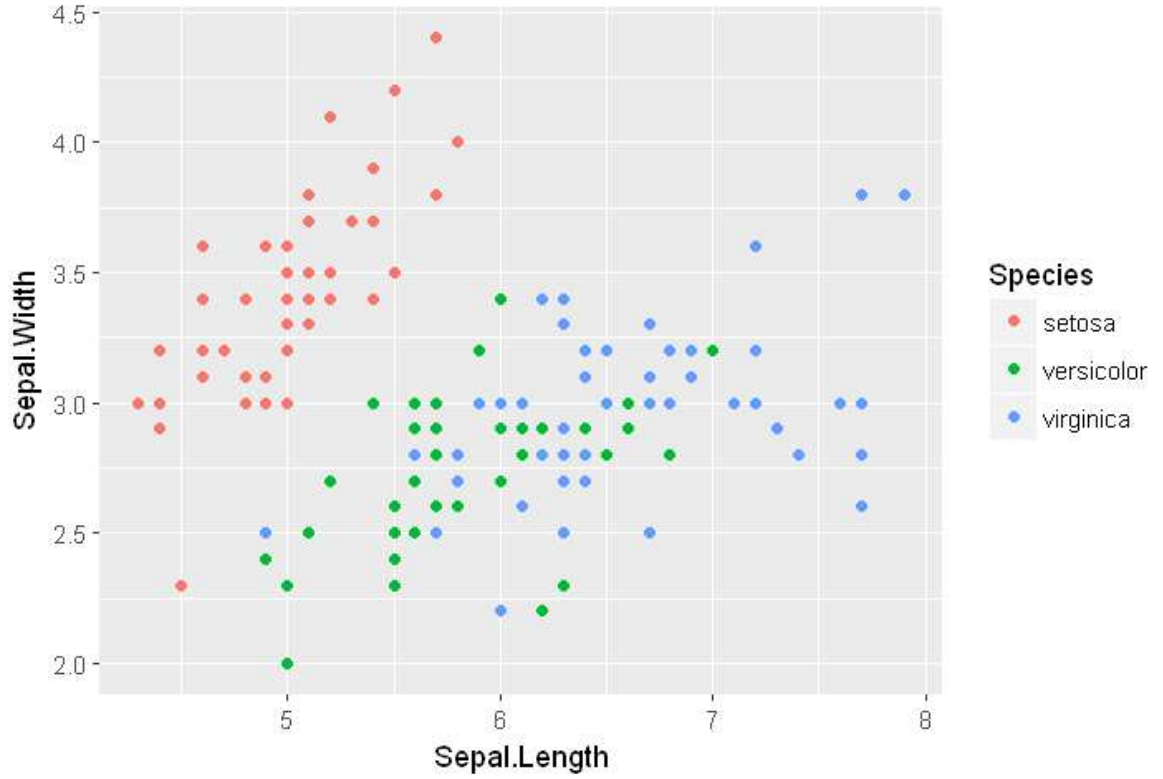
In [17]:
```
ggplot(data=dt.iris.melt, aes(x=variable, y=value, color=Species)) + geom_boxplot() + fac
et_wrap(~variable, ncol=4, scales = "free")
```



In [19]:
```
options(repr.plot.width=6, repr.plot.height=4)
```

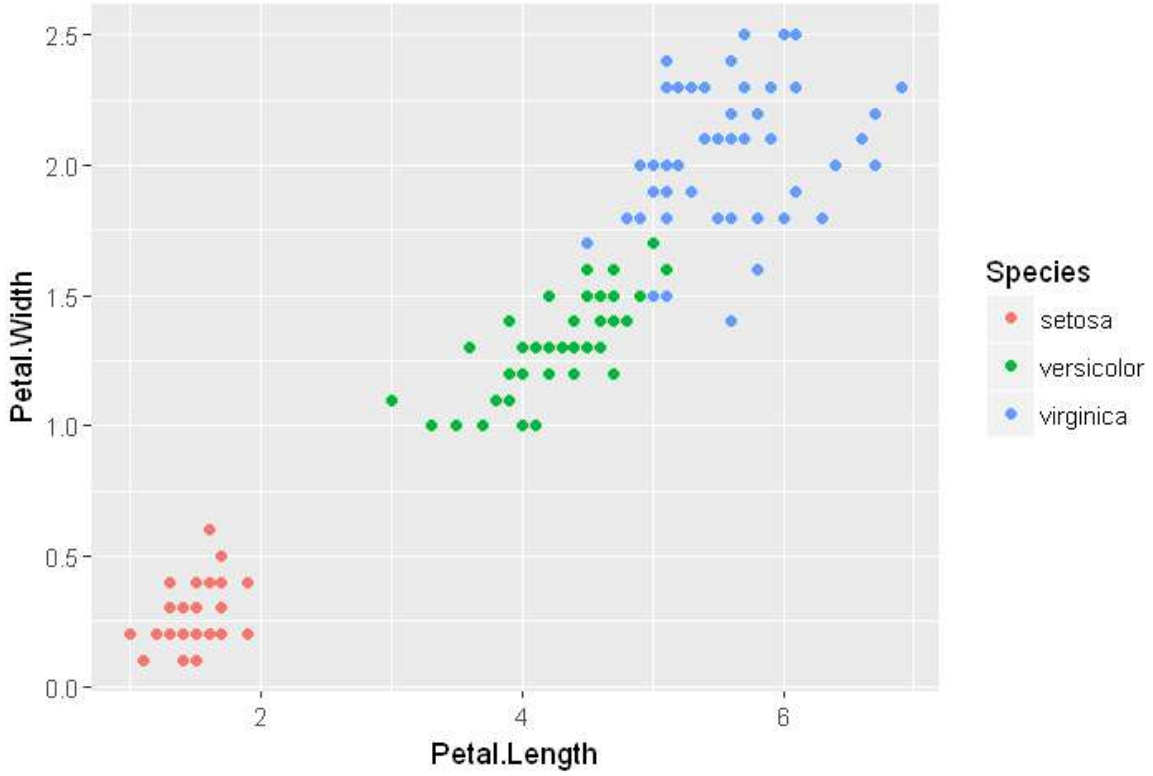
- Interação dos atributos *Sepal.Length* and *Sepal.Width* por *Species*

In [23]:
```
ggplot(data=dt.iris, aes(x=Sepal.Length, y= Sepal.Width, color = Species)) + geom_point()
```



- Interação dos atributos *Petal.Length* and *Petal.Width* por *Species*

In [25]:
```
ggplot(data=dt.iris, aes(x=Petal.Length, y= Petal.Width, color = Species)) + geom_point()
```



Mineração de Dados

Objetivo: criar uma função que utilize os 4 atributos numéricos para determinar a espécie da flor.

Se conseguirmos encontrar uma boa função para determinar a espécie, podemos usar essa função para descobrir a espécie de novas flores.

In [26]:
```
library(RWeka)
library(partykit)
options(repr.plot.width=8, repr.plot.height=5)
```

```
Loading required package: grid
```

In [27]:
```
(model <- J48(data = dt.iris, Species ~. ) )
```

```
J48 pruned tree
------------------

Petal.Width <= 0.6: setosa (50.0)
Petal.Width > 0.6
|   Petal.Width <= 1.7
|   |   Petal.Length <= 4.9: versicolor (48.0/1.0)
|   |   Petal.Length > 4.9
|   |   |   Petal.Width <= 1.5: virginica (3.0)
|   |   |   Petal.Width > 1.5: versicolor (3.0/1.0)
|   Petal.Width > 1.7: virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9
```

Avaliação de Modelos

Verificar qual foi a precisão do modelo construído com os dados que temos

In [151]:
```
summary(model)
```

```
=== Summary ===

Correctly Classified Instances         147               98      %
Incorrectly Classified Instances         3                2      %
Kappa statistic                          0.97
Mean absolute error                      0.0233
Root mean squared error                  0.108
Relative absolute error                  5.2482 %
Root relative squared error             22.9089 %
Total Number of Instances              150

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = setosa
  0 49  1 |  b = versicolor
  0  2 48 |  c = virginica
```

Aprendizado Supervisionado e Não-Supervisionado

As técnicas para aprender modelos a partir de dados são classificadas de acordo com o tipo de aprendizado em:

- **Aprendizado Supervisionado**
- **Aprendizado Não-supervisinado**
- **Aprendizado Semi-Supervisionado**

Aprendizado Supervisionado

No aprendizado supervisionado os exemplos estão constituídos por atributos, onde um desses atributos é *classe*

O atributo *classe* permite orientar o processo de aprendizando para ajustar um modelo aos dados

Exemplo	Atributo 1	Atributo 2	...	Atributo m	Atributo classe
e_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$	y_1
e_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,m}$	y_2
...	...	...	...	...	...
e_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$	y_n

Neste aprendizado o objetivo consiste em encontrar uma função $f(X) = \hat{y}$, em que, dado os valores dos m atributos, estime y.

As tarefas de interesse do aprendizado supervisionado são preditivas e definidas por:

- **Classificação**: quando o valor da *classe* é um valor categórico (ou discreto)
- **Regressão**: quando o valor da *classe* é um valor numérico.

Aprendizado Não-superivsionado

No aprendizado não-supervisionado os exemplos estão constituídos por atributos, mas não há um atrbiuto *classe*

Isso indica que não tem como orientar o processo de aprendizado tentando acertar algum valor de atributo específico

Exemplo	Atributo 1	Atributo 2	...	Atributo m
e_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$
e_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,m}$
...	...	...	...	...
e_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$

Neste tipo de aprendizado tentamos encontrar grupos de dados que apresentem comportamento similar

As tarefas de interesse dentro do aprendizado não-supervisionado são:

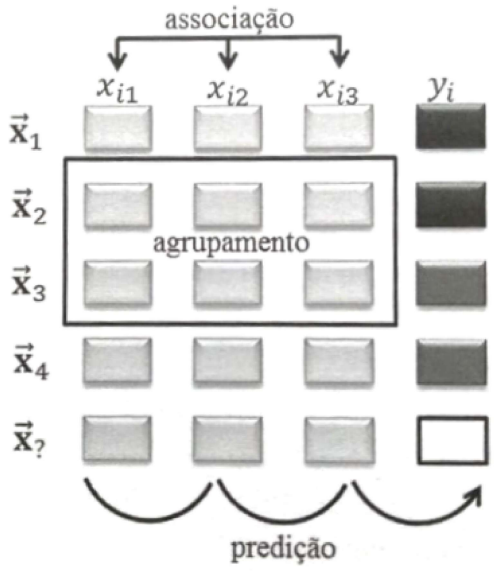
- **Agrupamentos**: agrupar exemplos de acordo com algum critério de distância ou similaridade
- **Extração de regras de associação**: encontrar relaçoes entre atributos e os valores dos atributos

Tarefas de Mineração de Dados

As principais tarefas de mineração são:

- Predição (seja classificação ou regressão)
- Agrupamento
- Associação

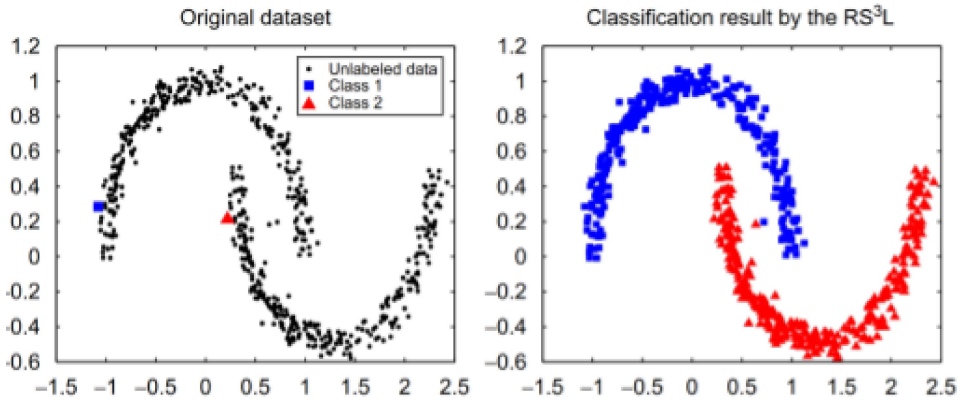
A seguinte figura temos a relação dessas tarefas em um conjunto de dados convencional:



Fonte: Silva, L. A., Peres, S. M., & Boscarioli, C. (2017). **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil

Aprendizado Semi-Supervisionado

- No aprendizado semi-supervisionado apenas alguns exemplos do conjunto de dados possuem o valor do atributo *classe*.
- Muitas vezes é muito caro classificar elementos. Alguns exemplos:
  - Transcrever segmentos de áudio
  - Identificar observações de experimentos físicos
  - Identificar os rostros de fotos do Facebook manualmente
- Nestas situações, com algumas instâncias rotuladas (exemplos classificados) é possível inferir os rótulos e classes de outras instâncias.



Fonte: Wang, F., & Zhang, C. (2007). **Robust self-tuning semi-supervised learning**. *Neurocomputing*, 70(16-18), 2931-2939.

Paradigmas de Aprendizado

Alguns autores chamam de paradigmas os aprendizados supervisionado e não-supervisionado (CASTRO e FERRARI, 2016).

No entanto a literatura clássica divide os paradimgas de aprendizado da seguinte forma:

- Simbólico
- Baseado em Instância
- Estatístico
- Conexionista
- Genético

Uma ótima leitura sobre os pardigmas pode ser encontrada em:

- MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos sobre aprendizado de máquina**. Sistemas Inteligentes-Fundamentos e Aplicações, v. 1, n. 1, p. 32, 2003. Disponível em: http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf (http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf)