



MBA⁺

**ARTIFICIAL
INTELLIGENCE
& MACHINE
LEARNING**



MBA⁺

Plataformas Cognitivas

Prof.: MARCIO JUNIOR VIEIRA
Email: marcio@ambientelivre.com.br

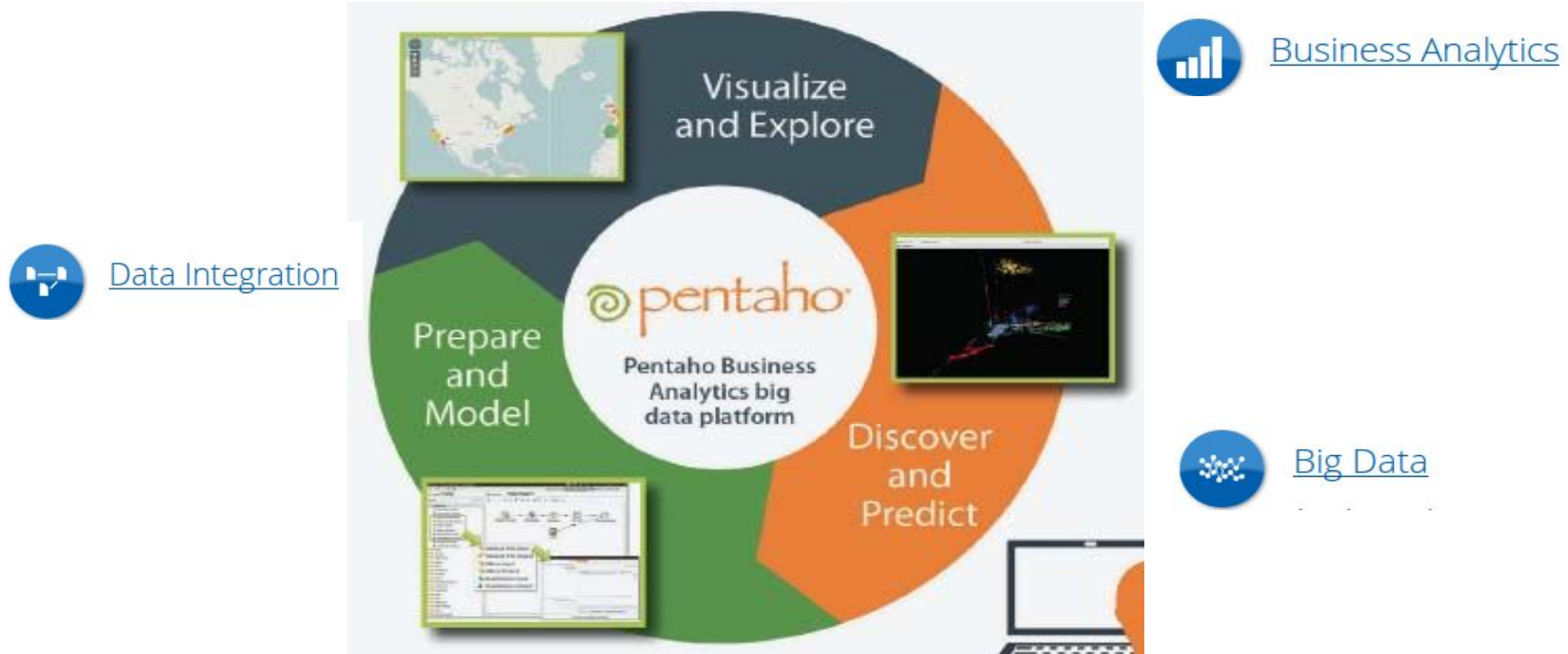
Weka

- Desenvolvido pela Universidade de Waikato (**W**aikato **E**nvironment for **K**nowledge **A**nalysis)
- Licença GPL
- Desenvolvido em Java
- Iniciado o desenvolvimento em 1993.
- O software foi adquirido pela Pentaho Corporation em 2016 (Hoje chamada de Hitachi Vantara).
- Site do projeto: <http://www.cs.waikato.ac.nz/ml/weka/>



3 Pilares do Pentaho

- Plataforma abrangente para integração de dados e Business Analytics.



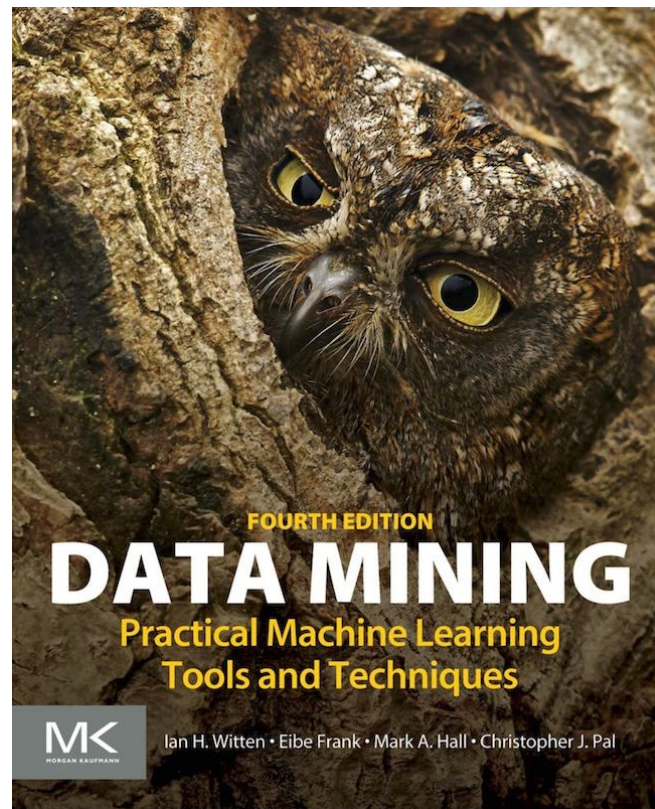
- Aprendizado de máquina
- Mineração de Dados
- Pré-processamento
- Classificação
- Regressão
- Agrupamento
- Regras de associação
- Atributo de seleção
- Experiências
- Workflow
- Visualização



- Líder do projeto Weka / PDM.
- 15 anos de experiência como pesquisador acadêmico em ciências da computação.
- Diversas publicações em conferências de aprendizado de máquina, mineração de dados e revistas.
- Se formou no pós-doutorado da Universidade de Waikato, na Nova Zelândia.
- Blog: markahall.blogspot.com
-



- Stable – Weka 3.8.2
- Developer – Weka 3.9
- Book Data Mining – Weka 3.6



- Repositório no SourceForge - <https://sourceforge.net/projects>
- Site Oficial <http://www.cs.waikato.ac.nz/ml/weka>
- Licença: GNU General Public License version 3.0 (GPLv3)





Software Libre

Open Source

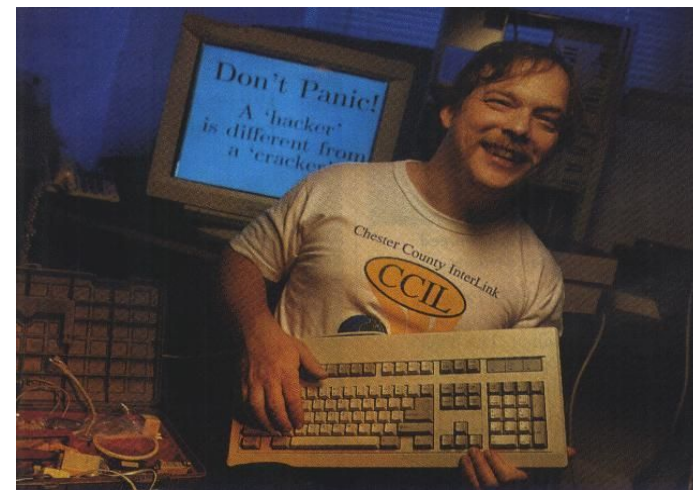
- **"Software Livre"** se refere à liberdade dos usuários executarem, copiarem, distribuírem, estudarem, modificarem e aperfeiçoarem o software. São **4 tipos de liberdade**, para os usuários do software:
- **1.** A liberdade de executar o programa, para qualquer propósito.
- **2.** A liberdade de estudar como o programa funciona, e adaptá-lo para as suas necessidades. Acesso ao código-fonte é um pré-requisito para esta liberdade.
- **3.** A liberdade de redistribuir cópias de modo que você possa ajudar ao seu próximo.
- **4.** A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.

Stallman



Dr. Richard

- Criado pela OSI (Open Source Initiative)
- Não refere-se a software também conhecido por software livre.
- **Qualquer licença de software livre é também uma licença de código aberto (Open Source)**
- **Mas o contrário nem sempre é verdade**
- Criado por Eric Raymond entre outros fundadores da OSI.
- Raymond



Hacker Eric

Diferença: Free Software X Open Source

- 4 Lei da GPL
- **Obrigatoriedade no Free Software (GPL) opcional na Open Source**
A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.



X



- **Explorer:** Uso geral (Pré-processamento, clusterização Classificação, visualização)
- **Experimenter:** controle de treinamento (divisão do conjunto teste/treinamento, cross-validation)
- **KnowledgeFlow:** Tarefas de ETL como fluxo de dados
- **Workbench:** GUI antiga.
- **Simple CLI:** Console para uso por linha de comando

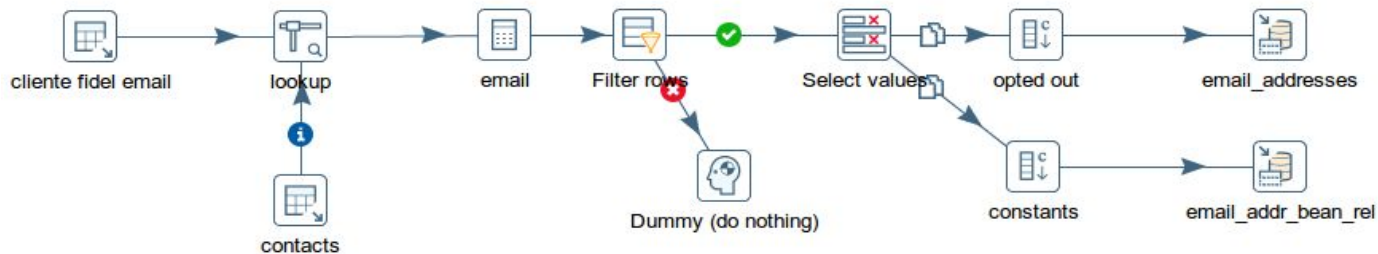
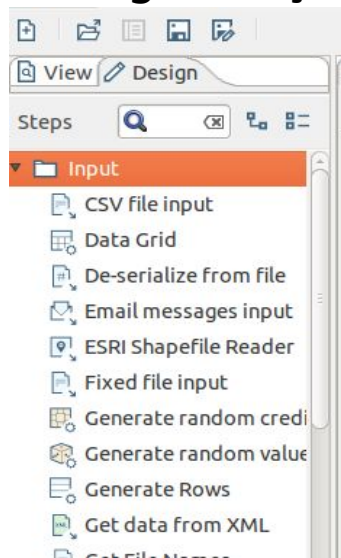


- Suporte vários formatos de dados como : Texto, URL, Database(JDBC), e geração automática
- ARFF é o mais comum!
- Para utilizar as bibliotecas da ferramenta Weka é necessário adaptar os dados a um formato de base de dados próprio do Weka chamado ARFF (Attribute-Relation File Format).
- E antes de aplicar os dados aos algoritmos do pacote Weka estes devem ser convertidos para o formato ARFF na grande maioria
- O ARFF que basicamente é um arquivo no formato ASCII.
-
- Documentação sobre o ARFF: <http://weka.wikispaces.com/arff>

- @relation
- @attribute
- @data

```
1 @relation tempo
2
3 @attribute situacaotempo {ensolarado, nublado, chuvoso}
4 @attribute temperatura real
5 @attribute humidade real
6 @attribute ventoso {VERDADEIRO, FALSO}
7 @attribute jogo {SIM, NAO}
8
9 @data
10 ensolarado,85,85,FALSO,NAO
11 ensolarado,80,90,VERDADEIRO,NAO
12 nublado,83,86,FALSO,SIM
13 chuvoso,70,96,FALSO,SIM
14 chuvoso,68,80,FALSO,SIM
15 chuvoso,65,70,VERDADEIRO,NAO
16 nublado,64,65,VERDADEIRO,SIM
17 ensolarado,72,95,FALSO,NAO
18 ensolarado,69,70,FALSO,SIM
19 chuvoso,75,80,FALSO,SIM
20 ensolarado,75,70,VERDADEIRO,SIM
21 nublado,72,90,VERDADEIRO,SIM
22 nublado,81,75,FALSO,SIM
23 chuvoso,71,91,VERDADEIRO,NAO
```

- Processa em Paralelo (também roda em Cluster Spark)
- Acessar dados diretamente (se necessário sem DW)
- Permite publicar dados diretamente em Reports, Ad-Hoc Reports e Dashboards.
- “Programação e Fluxo Visual” com aproximadamente 350 steps



Funcionalidades Tradicionais

- Usadas em projetos de data warehouse

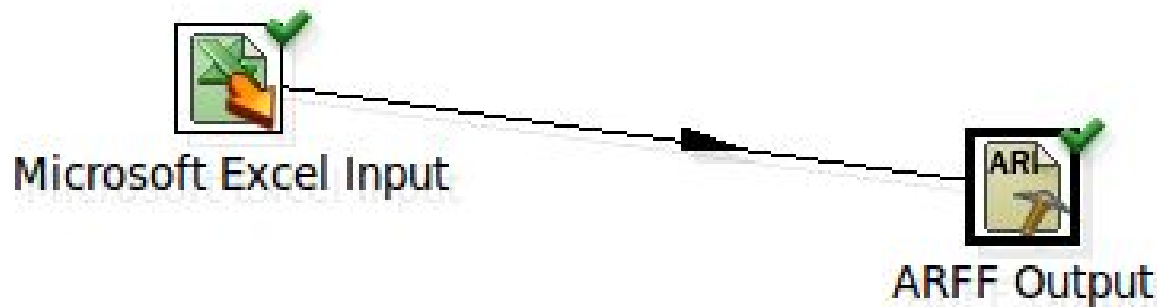
Funcionalidades Adicionais

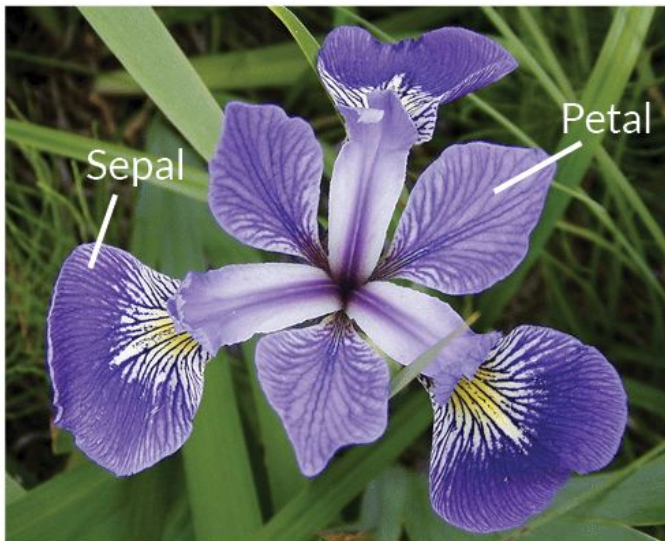
- Migração de dados entre aplicações/banco de dados
- Exportar dados de banco de dados para arquivos texto
- Carregar massivamente dados em banco de dados
- Data Cleansing – disciplina de qualidade/limpeza de dados de data warehouse
- Integração de aplicações.
- Gerenciamento de Filesystem (File management)
- **Geração de Arquivo ARFF**



Step ARFF

- Deve ser instalado o weka.jar no diretório lib do PDI.
- Deve ser instalado o plugins no Kettle.





Iris Versicolor

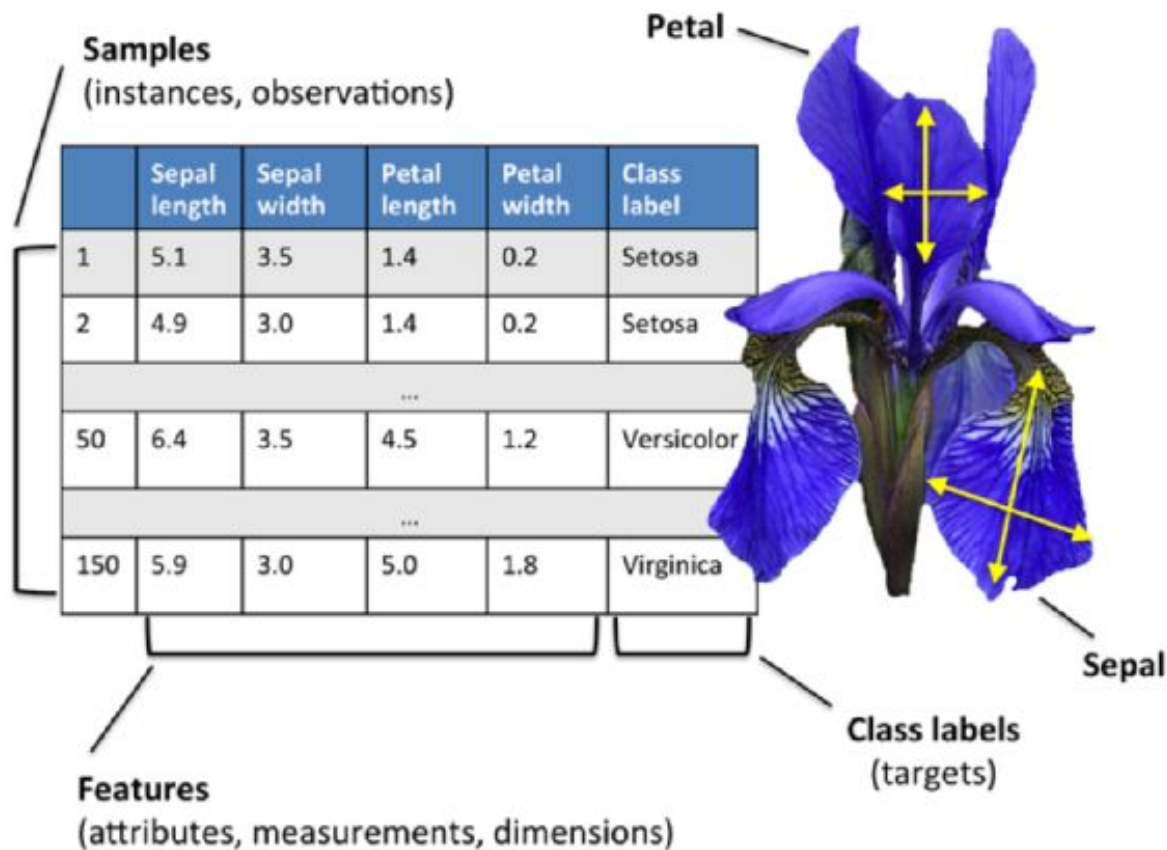


Iris Setosa



Iris Virginica

Supervisionado - Extração de Características



Aprendizagem Supervisionada

Fase 1



Extração de Características

Algoritmo de Aprendizagem

Modelo Preditivo



Supervisor

Fase 2



Extração de Características

Modelo Preditivo

Iris Versicolor

Data Sets

- Golf - Jogos.
- Tipos de Iris - Botânica.
- Supermercado - Negócios.
- Diabetes - Saúde

Vamos Executar este :)

- Classificação - J48 – Árvores de decisão
- Associação – Apriori



MBA⁺

Copyright © **2019** Prof. Marcio
Junior Vieira

Todos direitos reservados.
Reprodução ou divulgação total
ou parcial deste documento é
expressamente proibido sem o
consentimento formal, por
escrito, do Professor (autor).