



Encontros  
Pré-ConfOA  
2017



Encontros que antecedem a 8ª Conferência Luso-Brasileira de Acesso Aberto  
julho a setembro de 2017

# Plataforma de Ciência de Dados aplicada à Saúde

**Prof. Dr. Marcel Pedroso**  
Pesquisador em Saúde  
Pública

## CIÊNCIA DE DADOS - DEFINIÇÃO

Ciência de Dados é um conjunto de estratégias, ferramentas e técnicas que busca reunir equipes multidisciplinares formadas por pesquisadores com conhecimento substantivo do problema em análise - no nosso caso saúde pública - estatísticos, matemáticos e cientistas da computação. Trata-se de um campo de estudo promissor e destaca-se pela capacidade de auxiliar a descoberta de informação útil a partir de grandes bases de dados e a tomada de decisão orientada por dados

## PESQUISA CIENTÍFICA PARA O SUS

### Laboratórios de Pesquisa:



- Laboratório de Informação em Saúde (LIS)



- Laboratório de Informação Científica e Tecnológica em Saúde (LICTS)



- Laboratório de Comunicação e Saúde (LACES)

## INSTITUIÇÕES PARCEIRAS



Laboratório  
Nacional de  
Computação  
Científica



DEXL LAB  
EXTREME DATA LAB



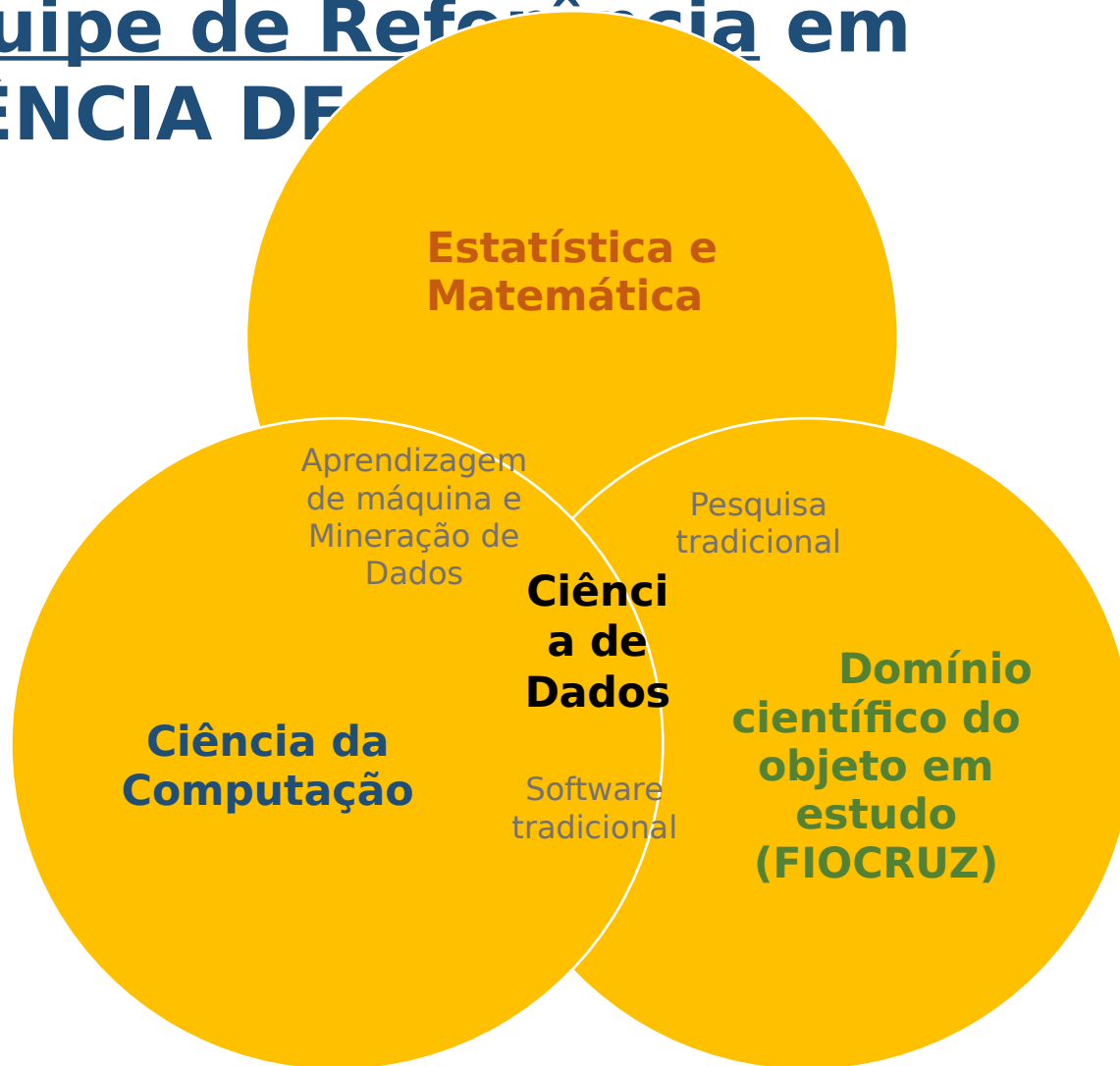
LABDAPS  
LABORATÓRIO DE BIG DATA E  
ANÁLISE PREDITIVA EM SAÚDE

EIC

Escola de Informática e Computação



# Equipe de Referência em **CIÊNCIA DE DADOS**

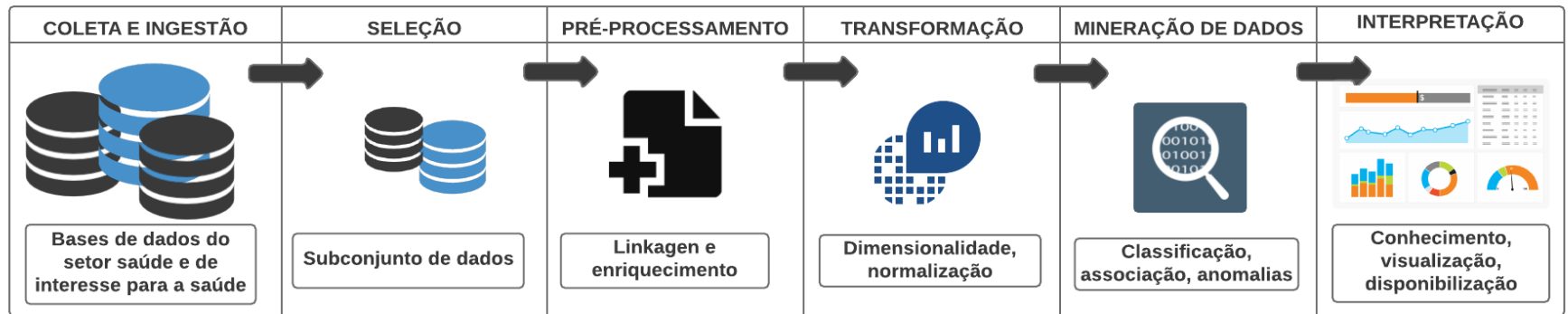


## OLUME E VARIEDADE (até 2015)

- ✓ **180 milhões** de autorizações de internações hospitalares (Datassur)
- ✓ **57 milhões** de registros de nascimentos (Datassur)
- ✓ **19 milhões** de declarações de óbito (Datassur)
- ✓ **24 bilhões** de registros de procedimentos ambulatoriais (Datassur)
- ✓ **2,5 bilhões** de registros de doses de vacinas (Datassur)
- ✓ **120 milhões** de questionários domiciliares dos Censos de 2000 e 2010 (IBGE)
- ✓ **milhões** de medições diárias realizadas pelo INPE e CEMADEN
  
- ✓ diariamente são criados **2,5 quintilhões** de dados, sendo que **3/4** deste conteúdo é produzido por internautas (dados interacionais) e não empresas ou instituições (dados transacionais)
- ✓ As **interações** com Google geram diariamente 24 petabytes (1 PB é o suficiente para armazenar, por exemplo, um vídeo em alta definição com duração de 13,3 anos)
- ✓ Os 1.3 bilhões de **usuários** do Facebook (68 milhões no Brasil) compartilham mais de 350 milhões de fotos e geram 500 terabytes por dia

# Cobertura de Conhecimento em Bases de Dados

## Knowledge Discovery in Databases - KDD



DADOS NÃO ESTRUTURADOS →

← DADOS ESTRUTURADOS

# PLATAFORMA DE CIÊNCIA DE DADOS como serviço (PaaS)



## COMPONENTES DA PLATAFORMA

### Indexação, Extração e Análise Visual

 Elasticsearch

 Kibana

### Mineração de Dados e Análise Preditiva

 R Studio

### Inovação e Aprendizagem Colaborativa

 Data Science Lab



# COMPONENTE ANÁLISE VISU

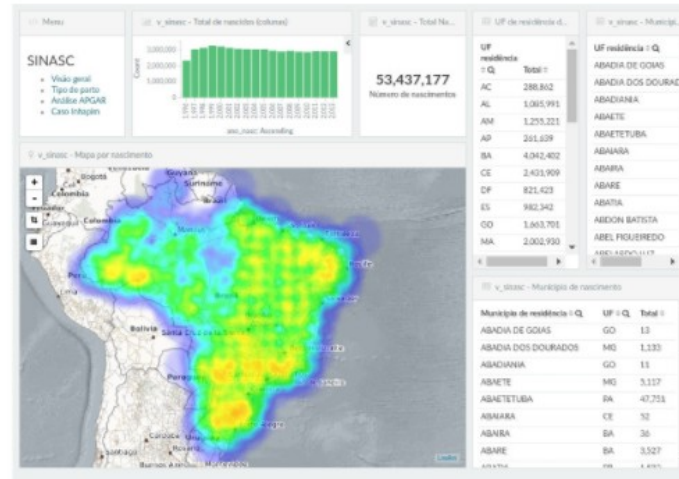


## Análise Visual

Indexação, extração e análise visual de grandes quantidades de dados do setor saúde e seus determinantes socioambientais

### Acesso qualificado aos dados da Plataforma

- Indexação e disponibilização de grandes bases de dados
- Extração de subconjuntos de dados de interesse para os pesquisadores
- Análise visual de situações de saúde
- Processamento distribuído e escalável



[Acesse a plataforma](#)

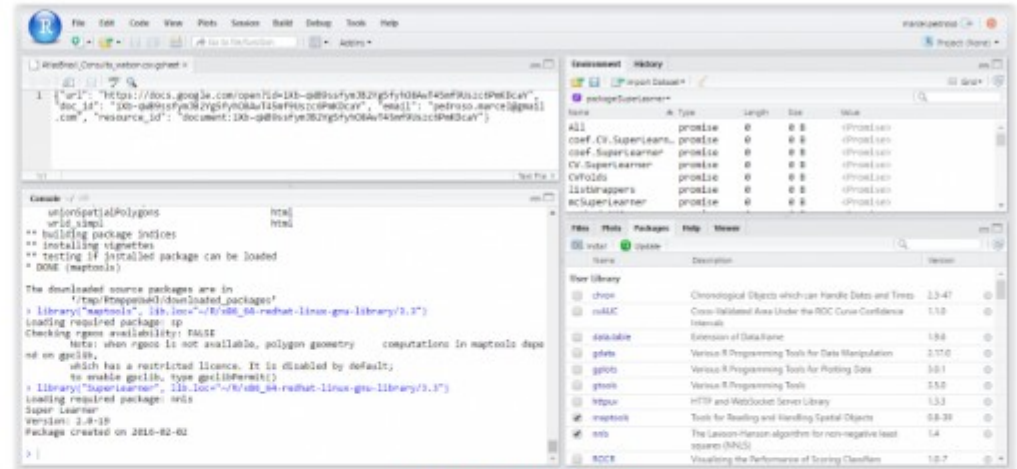
# COMPONENTE MINERAÇÃO DE DADOS



## Mineração de Dados e Análise Preditiva Conexão aos dados da Plataforma via R Studio Server

### Conexão e análise dos dados na Plataforma utilizando os pacotes R

- Conexão dos pacotes R sem necessidade de movimentação dos dados
- Utilização da infraestrutura de processamento e armazenamento da Plataforma
- Processamento dos dados por meio de escalonamento, replicação e indexação via Elasticsearch



Acesse a plataforma

# COMPONENTE DATA SCIENCE



## Data Science Lab

### Inovação e Aprendizagem Colaborativa



**Kaggle**  
Plataforma de aprendizagem e competição para cientistas de dados.



**Jupyter Python**  
Plataforma de ciência de dados para compartilhamento de códigos em mais de 40 linguagens de programação.



**Curso online de Ciência de Dados**  
Programa de cursos da Universidade Johns Hopkins na Coursera.



**Curso online de Ciência de Dados**  
Aprenda a programar em Python de um modo profissional e divertido.



**Big Data University - IBM**  
Cursos gratuitos em português sobre Big Data.



**GitHub Educação**  
Codificação - Cursos e ferramentas grátis para estudantes.



**Codecademy**  
Aprenda a programar de maneira interativa e gratuita.



**Social Data Analytics**  
Mineração de texto e análise de sentimento por meio de redes sociais pela equipe IBM Research - Brasil.



**DataCamp**  
Aprenda Python e R na prática em plataforma inovadora de aprendizagem ciência de dados e machine learning.



**HARVARD UNIVERSITY**  
**The Harvard Data Science Initiative**  
Programa de pesquisa e ensino interdisciplinar que reunirá esforços em toda a Universidade de Harvard para o fomento de parcerias visando o fomento de iniciativas em Ciência de Dados.

## HARDWARE

- **2 servidores para de gestão do cluster**
- 4 Processadores Intel Xeon E5-2630 v3 com 8 Núcleos e 2.4GHz cada
- 256 GB memória RDIMM total (126 GB por servidor)
- **4 servidores para armazenamento de dados**
- 40 Terabyte de armazenamento
- 8 Processadores Intel Xeon E5-2630 v3 com 8 Núcleos e 2.4GHz cada
- 64 GB memória RDIMM total (16 GB por servidor)
- **10 Gigabit conexão internet**



## CAPACITAÇÃO em CIÊNCIA DE DADOS

✓ EQUIPE DE REFERÊNCIA

**ICICT - LNCC - USP - CEFET - MS - UnB**

Plataforma online de capacitação em Ciência de Dados **DataCamp**



✓ DISCIPLINA e ATUALIZAÇÃO - ICICT

**Ciência de Dados aplicada à Saúde**

Curso presencial (120 horas) TCC utilizando a Plataforma

## GRUPO DE PESQUISA NO CNPq



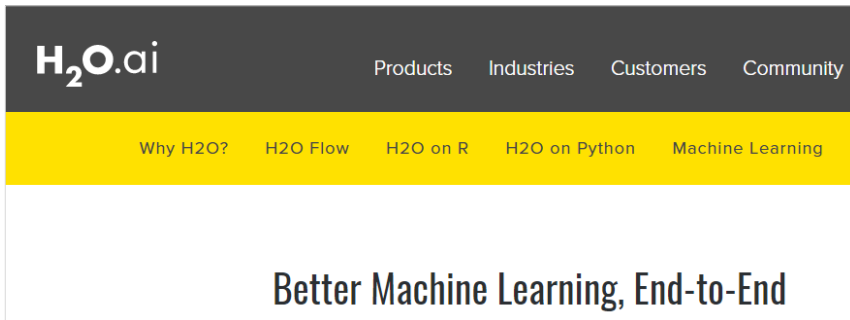
The screenshot shows the CNPq website interface for a research group. The header includes the CNPq logo and the 'Diretório dos Grupos de Pesquisa no Brasil Lattes' logo. The main content area displays the group's name, 'Ciência de Dados aplicada à Saúde', and a URL for the mirror page. A left sidebar contains a menu with options like 'Identificação', 'Endereço / Contato', and 'Equipamentos e Softwares'. The main content area shows the 'Identificação' section with details such as 'Situação do grupo: Certificado', 'Ano de formação: 2014', 'Data da Situação: 11/12/2014 09:53', 'Data do último envio: 18/05/2016 14:36', 'Lider(es) do grupo: Christovam Barcellos and Marcel de Moraes Pedroso', 'Área predominante: Ciências da Saúde, Saúde Coletiva', 'Instituição do grupo: Fundação Oswaldo Cruz - FIOCRUZ', and 'Unidade: Centro de Informação Científica e Tecnológica'. A 'CERTIFICADO PELA INSTITUIÇÃO' seal is also visible.

<http://dgp.cnpq.br/dgp/espelhogrupo/4230691756969719>

## PARAS PARCERIAS? (Instituições e pesquisadores)

[marcel.pedroso@icict.fiocruz.br](mailto:marcel.pedroso@icict.fiocruz.br)

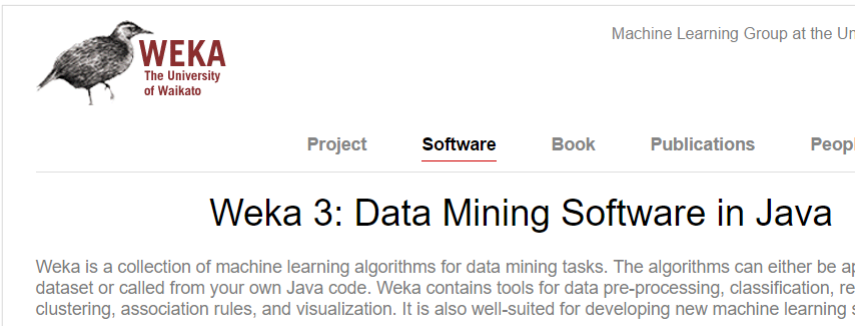
# 5 PASSOS - MINERAÇÃO DE DADOS E MACHINE LEARNING VISUAL




**H2O.ai** Products Industries Customers Community

Why H2O? H2O Flow H2O on R H2O on Python Machine Learning

## Better Machine Learning, End-to-End



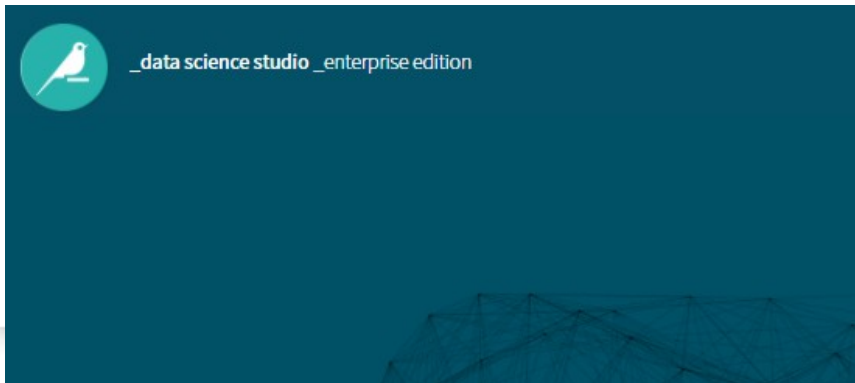
 **WEKA**  
The University of Waikato


Machine Learning Group at the University of Waikato

Project Software Book Publications People

## Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning software.



 **\_data science studio\_** enterprise edition

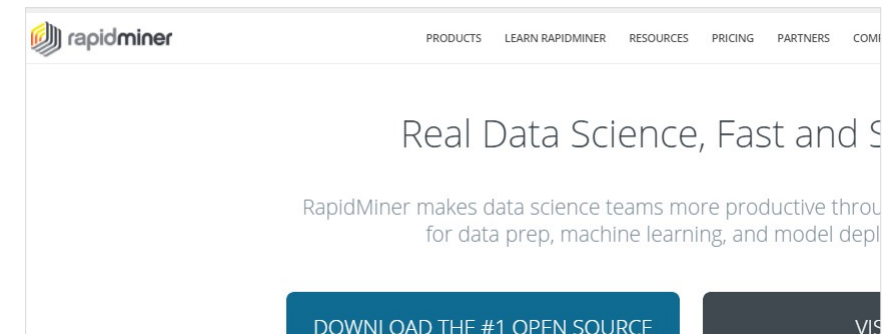



Open for Innovation <sup>®</sup>  
**KNIME** PRODUCTS / SOLUTIONS / LEARNING / PARTNERS / COMMUNITY / ABOUT

## Open for Innovation

Navigate complex data with the agility and freedom that only an open platform can bring

[Learn More](#)

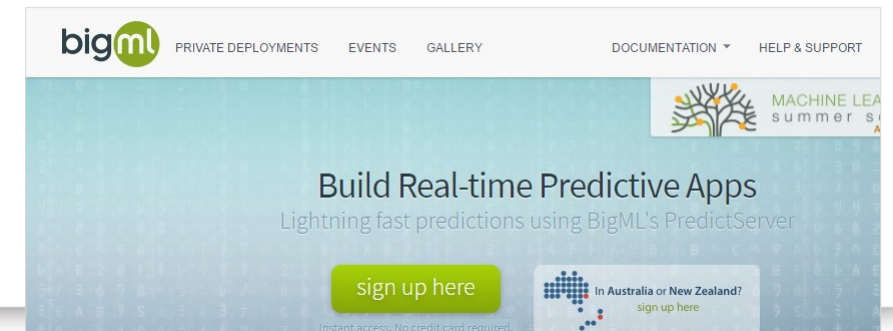



 **rapidminer** PRODUCTS LEARN RAPIDMINER RESOURCES PRICING PARTNERS COMMUNITY

## Real Data Science, Fast and Simple

RapidMiner makes data science teams more productive through automation for data prep, machine learning, and model deployment.

[DOWNLOAD THE #1 OPEN SOURCE](#)



 **bigml** PRIVATE DEPLOYMENTS EVENTS GALLERY DOCUMENTATION HELP & SUPPORT

MACHINE LEARNING summer school

## Build Real-time Predictive Apps

Lightning fast predictions using BigML's PredictServer

[sign up here](#)

Instant access. No credit card required.

In Australia or New Zealand?  
[sign up here](#)

# CRONOGRAMA

Meta	Fase	mês 1	mês 2	mês 3	mês 4	mês 5	mês 6	mês 7	mês 8	mês 9	mês 10	mês 11	mês 12	mês 13	mês 14	mês 15	mês 16	mês 17	mês 18	mês 19	mês 20	mês 21	mês 22	mês 23	mês 24
1	<b>FONTES DE DADOS</b>																								
1.1	Pesquisar e avaliar ferramentas tecnológicas e estratégias para o desenvolvimento da infraestrutura																								
1.2	Identificar, coletar, normalizar, armazenar e conectar as bases de dados de interesse																								
2	<b>GESTÃO E ARMAZENAMENTO</b>																								
2.1	Adquirir equipamentos para a Plataforma																								
2.2	Desenvolver, testar e homologar a Plataforma																								
3	<b>ANÁLISE E VISUALIZAÇÃO</b>																								
3.1	Análise visual, análise multicritério, mineração de dados e análise preditiva de Big Data em Saúde																								
3.2	Capacitar equipe de referência em Computação Científica e Big Data																								

## PRINCIPAIS PARCEIROS



Ministério da Saúde





# Institute for Scientific and Technological Communication and Information on Health

[www.facebook.com/fiocruz.icict](https://www.facebook.com/fiocruz.icict)

[twitter.com/@Icict\\_fiocruz](https://twitter.com/Icict_fiocruz)

[www.youtube.com/videosaudefio](https://www.youtube.com/videosaudefio)

## [www.icict.fiocruz.br](http://www.icict.fiocruz.br)