Plataforma de Ciência de Dados aplicada à Saúde

Prof. Dr. Marcel Pedroso
Pesquisador em Saúde Pública

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

SUS

Ministério da Saúde

Governo Federal

IÊNCIA 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

ICICT

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)

ISTITUIÇÕES PARCEIRAS

Laboratório Nacional de Computação Científica

Universidade de São Paulo – Faculdade de Saúde Pública

CEFET/RJ

DEXL LAB
EXTREME DATA LAB

LABDAPS
Laboratório de Big Data e Análise Preditiva em Saúde

EIC
Escola de Informática e Computação

CEAG
Centro de Estudos Avançados de Governo e Administração Pública

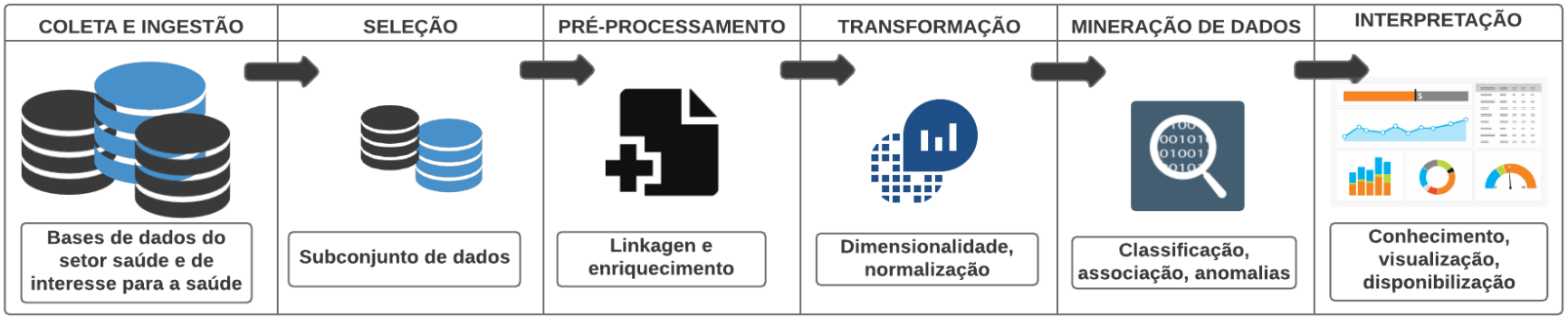
Equipe de Referência em CIÊNCIA DE

OLUME E VARIEDADE (até 2015)

- ✓ **180 milhões** de autorizações de internações hospitalares (Datasus)
- ✓ **57 milhões** de registros de nascimentos (Datasus)
- ✓ **19 milhões** de declarações de óbito (Datasus)
- ✓ **24 bilhões** de registros de procedimentos ambulatoriais (Datasus)
- ✓ **2,5 bilhões** de registros de doses de vacinas (Datasus)
- ✓ **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

coberta de Conhecimento em Bases de Dados
wledge Discovery in Databases – KDD



TAFORMA DE CIÊNCIA DE DADOS como serviço (PaaS)

COMPONENTES DA PLATAFORMA

Indexação, Extração e Análise Visual

Elasticsearch

Kibana

Mineração de Dados e Análise Preditiva

RStudio

Inovação e Aprendizagem Colaborativa

Data Science Lab

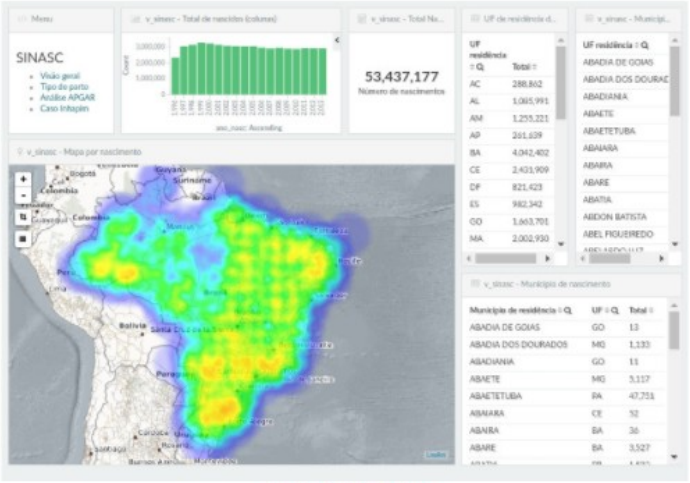
COMPONENTE ANÁLISE VISU

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

Elasticsearch

Kibana

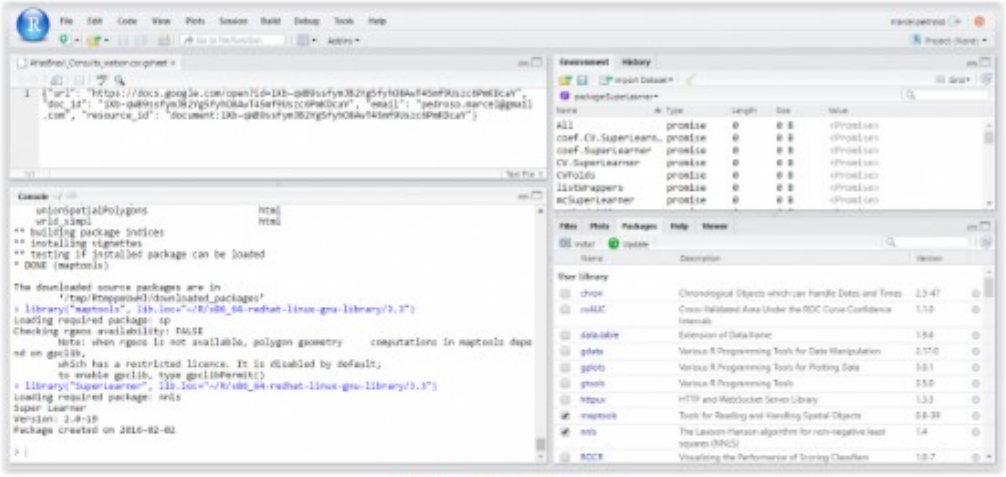
COMPONENTE MINERAÇÃO DE DADOS

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



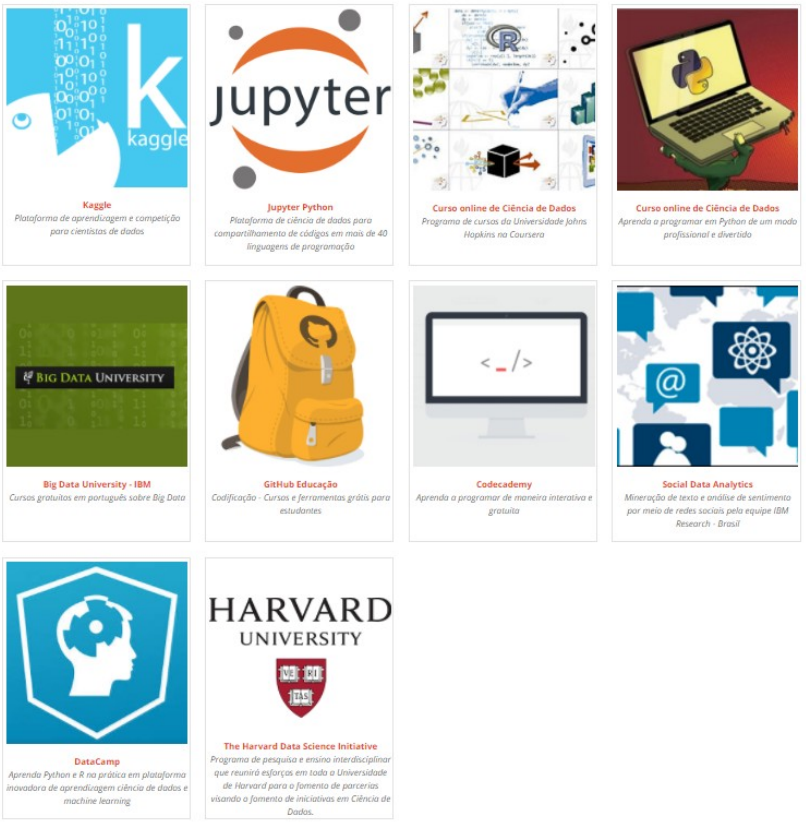
Acesse a plataforma

RStudio

COMPONENTE DATA SCIENCE

Data Science Lab

Inovação e Aprendizagem Colaborativa



HARDWARE

Laboratório Nacional de Computação Científica

- **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

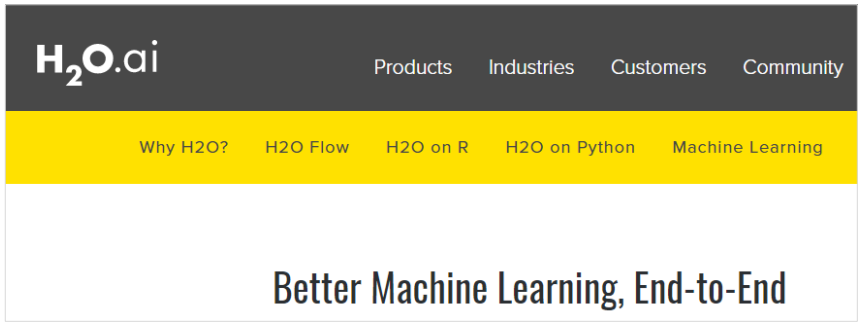
GRUPO DE PESQUISA NO CNPq

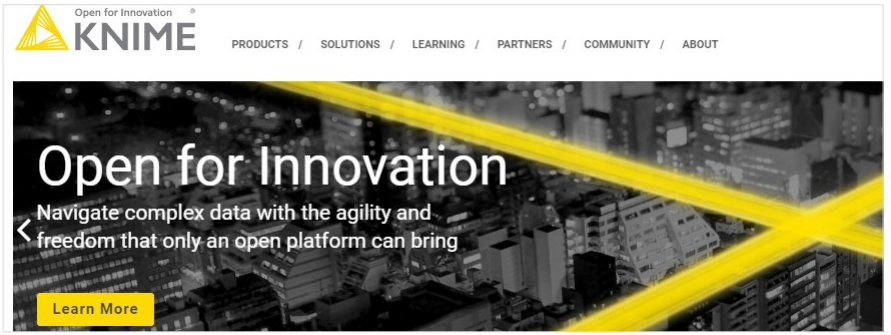
http://dgp.cnpq.br/dgp/espelhogrupo/42306917569697 19

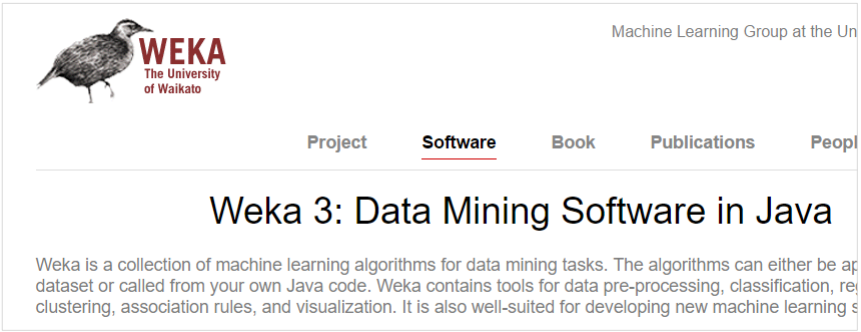
TRAS PARCERIAS? (Instituições e pesquisadores)

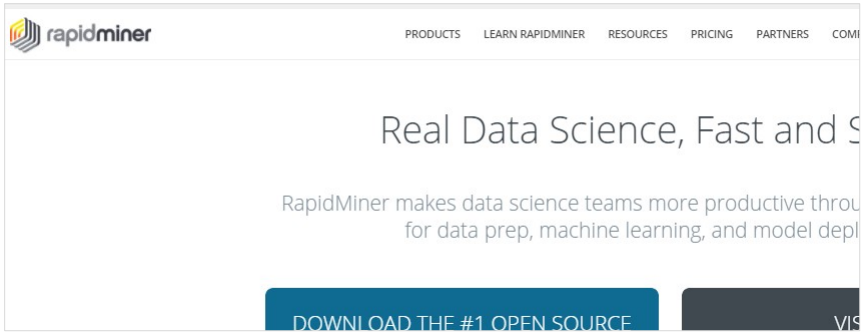
marcel.pedroso@icict.fiocruz.br

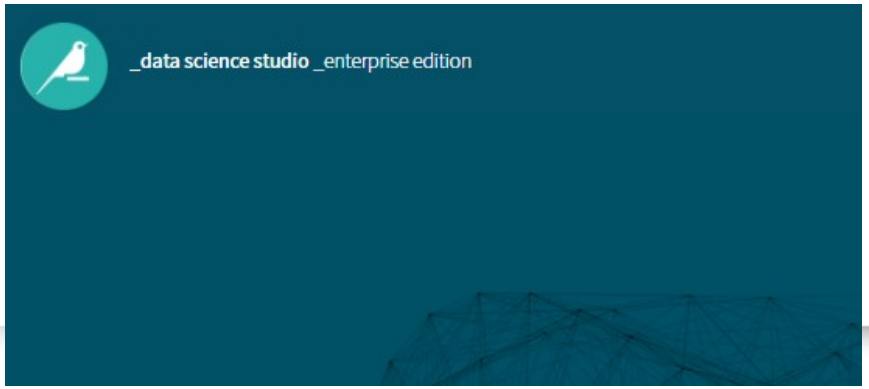
XIMOS PASSOS - MINERAÇÃO DE DADOS E MACHINE LEARNING VISUAL

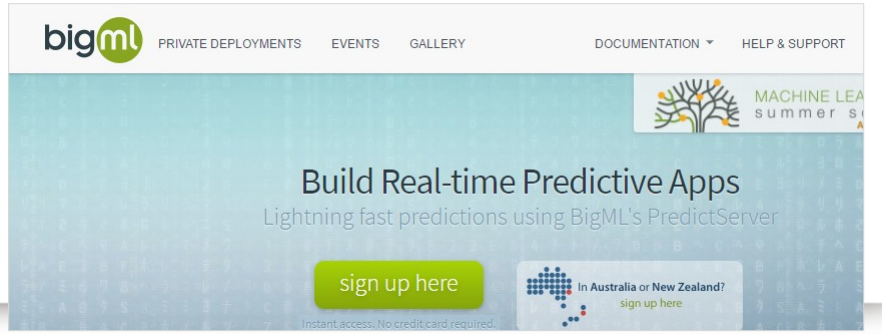












CRONOGRAMA

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	FONTES DE DADOS																								
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	GESTÃO E ARMAZENAMENTO																								
2.1	Adquirir equipamentos para a Plataforma																								
2.2	Desenvolver, testar e homologar a Plataforma																								
3	ANÁLISE E VISUALIZAÇÃO																								
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			Minicursos								Seminário				Minicursos								Seminário	

PRINCIPAIS PARCEIROS

Laboratório Nacional de Computação Científica

CEAG Centro de Estudos Avançados de Governo e Administração Pública

Universidade de São Paulo – Faculdade de Saúde Pública

CEFET/RJ

SUS Sistema Único de Saúde

Ministério da Saúde

DELL EMC

IBM

INPE

Cemaden Centro Nacional de Monitoramento e Alertas de Desastres Naturais

Institute for Scientific and Technological Communication and Information on Health

www.facebook.com/fiocruz.icict

twitter.com/@Icict_fiocruz

www.youtube.com/videosaudefio

www.icict.fiocruz.br

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

SUS

Ministério da Saúde

Governo Federal