



# Ciência de Dados

## Estado da Arte e Oportunidades para a CGU

Rodrigo Dewes (SCC/DIE)

[rodrigo.dewes@cgu.gov.br](mailto:rodrigo.dewes@cgu.gov.br)

Rodrigo Pedatella (SFC/DAE)

[rodrigo.pedatella@cgu.gov.br](mailto:rodrigo.pedatella@cgu.gov.br)



## *Roteiro*

1. Sobre ciência de dados
2. Sobre o KDD
3. O evento
4. Os tutoriais *hands-on*
5. Lições aprendidas
6. Oportunidades para a CGU

# Ciência de dados – o que é?



- **Estudo** de dados para possibilitar:
  - Extração
  - Análise
  - Visualização
  - Gerenciamento
  - Armazenamento
- Ciência multidisciplinar:
  - Estatística
  - Matemática
  - Ciência da computação
- **Grandes** massas de dados:
  - Não restrito, mas vantajoso
- Machine learning:
  - Automatização de decisões

# Um exemplo...



# Ciência de dados - o assunto da moda

Por que todo mundo está falando sobre isso?

1. Versatilidade (numerosos campos de atuação)
2. Enriquecimento de dados
3. Automação de tarefas redundantes
4. Aprendizagem de máquina (produtos melhores)
5. Ciência de dados na área de saúde





# O que é o KDD?

- Mais importante evento sobre ciência de dados do mundo.
- Ocorre anualmente, sob organização da ACM.
- Reúne os maiores pesquisadores do **meio acadêmico, indústria e governo** especializados em *data science, data mining, knowledge discovery, large-scale data analytics e big data*.
- Objetivo:
  - Compartilhar ideias, resultados de pesquisas e experiências.





# KDD 2019 em números

- 5 dias de atividades (04/08 a 08/08).
- 3 mil participantes.
- 2 centros de convenções:
  - Dena'ina Convention Center.
  - William Egan Convention Center.
- 2 keynote speakers e 1 plenary keynote panel.
  - Mais importantes cientistas de dados do mundo compartilham seu conhecimento para o avanço da aplicação de ciência de dados.
- 29 lecture-style tutorials.
  - Apresentações de cunho acadêmico aprofundadas sobre as tendências das pesquisas científicas.
- 34 workshops.
  - Discussões de ideias inovadoras com representantes do meio acadêmico, da indústria e do governo.
- 15 hands-on tutorials.
  - Demonstrações práticas de problemas do mundo real resolvidos com a aplicação de ciência de dados.
- 174 Research Track Papers.
- 148 Applied Data Science Track Papers.



# KDD - ferramenta de comunicação

11:43 88%  
Event Home

**KDD2019**

**25th ACM SIGKDD Conference**  
Aug 4 - 8, 2019  
Anchorage, AK

Quick Shortcuts

- Leaderboard
- Photos
- Speakers

Additional Resources

- Sponsors

Bronze **NAVER**

Home Agenda Attendees Community Messages

11:45 87%  
Tracks

Full Agenda My Agenda

sáb 3 dom 4 seg 5 ter 6 qua 7 **qui 8** sex 9 sáb 10

See 25 upcoming meet-ups in the community

8:00 AM

8:00 AM **KDD 2019 Registration**  
Location:Tikahtnu Foyer- Level 3, Dena'ina Center  
1 1 0

8:00 AM

8:00 AM **Keynote: Do Simpler Models Exist and How Can We Find Them?**  
Location:Tikahtnu Ballroom- Level 3, Dena'ina Cent...  
Speakers: Cynthia Rudin  
563 31 4

9:30 AM

9:30 AM **KDD Exhibit Hall**  
Location:Idlughet Hall- Street Level, Dena'ina Center  
28 0 0

9:30 AM

9:30 AM **Hands-On Tutorial: From Shallow to Deep Language Representations: ...**  
Location:Kahtnu 1- Level 2, Dena'ina Center  
236 13 1

Home Agenda Attendees Community Messages

11:50 87%  
Tracks

Full Agenda My Agenda

sáb 3 dom 4 seg 5 **ter 6** qua 7 qui 8 sex 9 sáb 10

See 25 upcoming meet-ups in the community

9:30 AM

9:30 AM **Hands-On Tutorial: Put Deep Learning to Work: A Practical Intro...**  
Location:Kahtnu- Level 2, Dena'ina Center  
285 24 0

12:00 PM

1:30 PM **Hands-On Tutorial: Cloud Based Data Science at the Speed of Thou...**  
Location:Kahtnu- Level 2, Dena'ina Center  
113 9 1

+ Add my own Activity

Home Agenda Attendees Community Messages

11:44 87%  
Community

Filter by: All Topics

**Organizer Announcements** Aug. 8  
Student Volunteers...stand and be recognized!  
We would love for all our amazing student volunteers to come to the Closing Ceremony at 4 pm in the Tikahtnu Ballroom at the Dena'ina C...  
17 messages (3 new)

**Meet-ups** Aug. 28  
SAN Diego animal park in Escondido Description: It's amazing - can't wait to go Location: Kdd 2020 one year from now Date: 08/28/2019 0...  
122 meet-ups (38 new meet-ups, 2495 new comments)

**Break the Ice!** Nov. 4  
Intro: Hello everyone! This is Haruyuki Sanuki. Looking forward to seeing you all!  
300 posts (300 new posts, 5 new comments) Follow

**How can we be useful for your research?** Oct. 1  
I created topic: How can we be useful for your research?. description: We are Ph.D. students and are building a produ...  
1 message (1 new) 1 Following Follow

**Job Openings** Aug. 30  
What job opportunities are you looking for? Hi everyone! I'm open to new opportunities. Feel free to check my experienc...  
227 posts (160 new posts, 17 new comments) Follow

**KDD 2020** Aug. 27  
Who are the committee members for KDD 2020, was it announced? Had to leave early to catch flight.  
2 messages (2 new) 1 Following Follow

ADD TOPIC OR SOCIAL GROUP

Home Agenda Attendees Community Messages

# KDD - o evento

## Domingo (04/08) - Tutorial Day:

- Abertura do evento.
- Dedicado a tutoriais em estilo de palestras.
- Eventos de 4h de duração.
- Apresentação de trabalhos acadêmicos e seus resultados.



# KDD - o evento

## Segunda (05/08) - Workshop Day:

- Dedicado a workshops em eixos temáticos:
  - Earth Day.
  - Deep Learning Day.
  - Health Day.
- Eventos de dia inteiro e eventos de 4h de duração.
- Apresentações temáticas com diferentes palestrantes.
- Discussões técnicas ao final de cada workshop.



# KDD - o evento

## Terça-feira (06/08) - Main Conference Day 1:

- *Keynote* de abertura com Peter Lee (Corporate Vice President, Microsoft Healthcare):
  - The Unreasonable Effectiveness, and Difficulty, of Data in Healthcare.
- Abertura do salão de exposições do KDD.
- Primeiras sessões de tutoriais *hands-on*:
  - Empresas de grande porte.
  - Demonstrações em tempo real.
  - Utilização de notebook próprio para conexão em plataformas de *data science*.
- Sessões de Research Track e Applied Data Science Oral Presentations:
  - Sessões menores, 2h de duração.



# KDD - o evento



## Quarta-feira (07/08) - Main Conference Day 2:

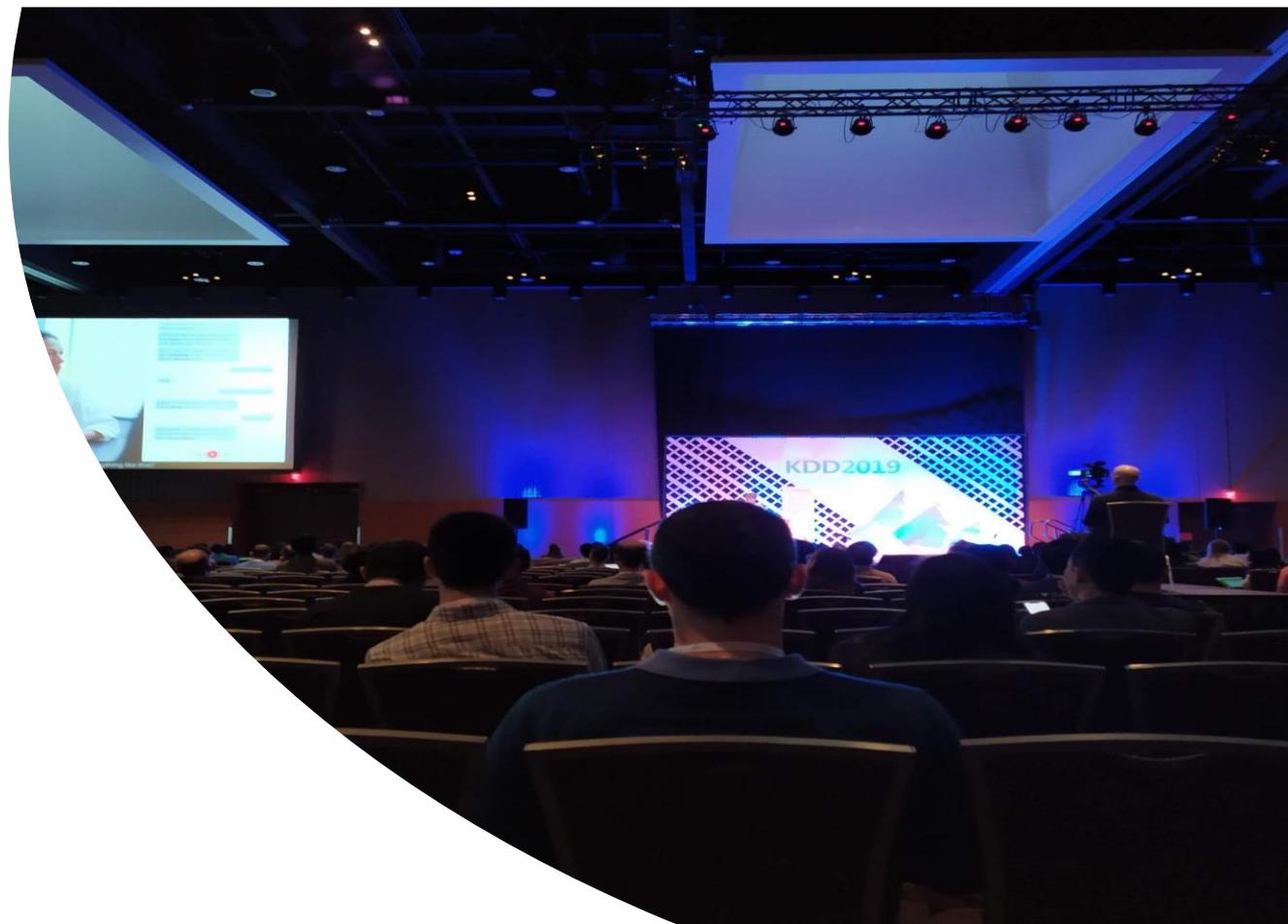
- Mais sessões de tutoriais *hands-on*.
- Mais sessões de Research Track e Applied Data Science Oral Presentations.
- KDD Project Showcase:
  - Challenges of Data Mining Projects (2 projetos)
  - Data & Society (6 projetos)
  - Demos (8 projetos)
  - Keynote (AI, Big Data, and the UN Sustainable Development Agenda)
  - Data for Earth Sensing (3 projetos)
  - Demos 2 (3 projetos)

# KDD - o evento

---

## Quinta-feira (08/08) - Main Conference Day 3:

- *Keynote* de encerramento com Cynthia Rudin (Duke University):
  - Do Simpler Models Exist and How Can We Find Them?
- Últimas sessões de tutoriais *hands-on*.
- Últimas sessões de Research Track Oral Presentations.
- KDD 2019 Closing Session.





# Tutoriais *hands-on*

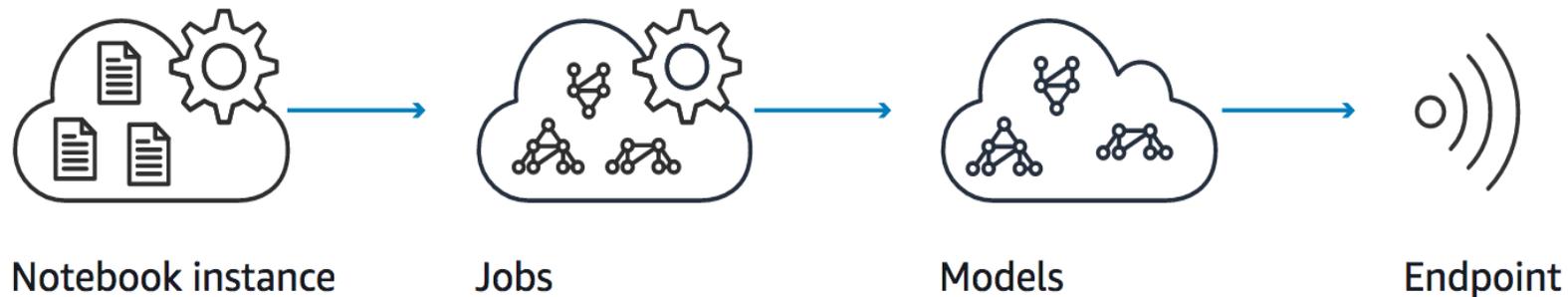
Participamos dos seguintes tutoriais:

- Amazon AWS – SAGEMaker
- Nvidia Rapids
- From Shallow to Deep Language Representations
- Democratizing & Accelerating AI through Automated Machine Learning
- Deep Learning at Scale on Databricks

<https://www.kdd.org/kdd2019/hands-on-tutorials>

# KDD2019 Amazon SageMaker Labs

- Plataforma para construção rápida de Notebooks de Machine Learning.
- Foram treinados algoritmos: Detecção de Objetos em Imagens e NLP (BERT).
- <https://github.com/awshlabs/kdd2019/blob/master/README.md>





# KDD2019 Amazon SageMaker Labs

The screenshot shows the AWS Management Console interface. On the left, under 'All services', the 'Machine Learning' category is expanded, and 'Amazon SageMaker' is highlighted with a red circle. The main area displays 'Helpful tips' and 'Explore AWS' sections.

The screenshot shows the 'Create notebook instance' configuration page. The 'Notebook instance type' dropdown is set to 'ml.p2.xlarge' and the 'Volume Size In GB' dropdown is set to '10', both highlighted with red boxes. The page includes sections for 'Notebook instance settings', 'IAM role', 'VPC', 'Lifecycle configuration', 'Encryption key', and 'Tags'.

# KDD2019 Amazon SageMaker Labs

- Classificou a imagem abaixo como pessoa.



# NVidia RAPIDS

- Uso de GPU em Machine Learning;
- Alto desempenho;
- Processamento Paralelo (multi-gpu);
- Processamento Distribuído (DASK);
- Fácil uso para quem já usa Pandas;
- Limitações: Memória da Placa de Vídeo.
- [https://github.com/rapidsai/notebooks-contrib/tree/master/conference\\_notebooks/KDD\\_2019](https://github.com/rapidsai/notebooks-contrib/tree/master/conference_notebooks/KDD_2019)



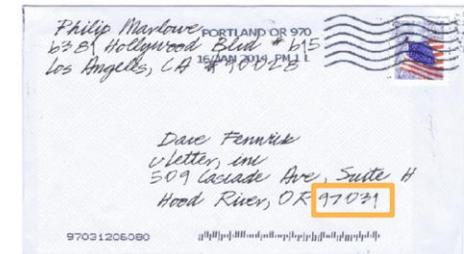
**RAPIDS**



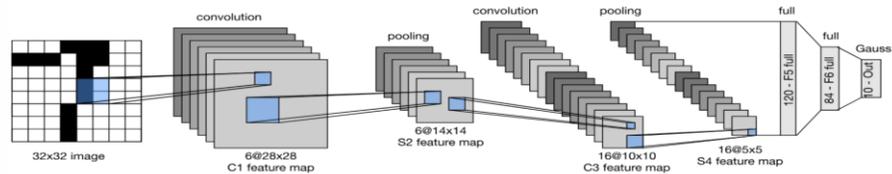
# From Shallow to Deep Language Representations

- Treino de processamento de linguagem natural – NLP;
- Comparativos entre: Numpy vs NDAarray (MXNet);
- <http://d2l.ai> (Dive into Deep Learning);
- Exemplos de Redes:

Handwritten Digit Recognition



## Convolutional Networks - LeNet



[d2l.ai](http://d2l.ai)



[d2l.ai](http://d2l.ai)





# From Shallow to Deep Language Representations

## NDArray (Linear Algebra on GPUs in Python)

- NumPy
  - Limited to CPUs
  - No automatic differentiation
  - **Blocking (returns only once computation is performed)**
- NDArray
  - Multiple CPUs / GPUs via device context
  - Distributed systems in the cloud
  - **Nonblocking C++ backend (no Python GIL)**
  - **Lazy evaluation**



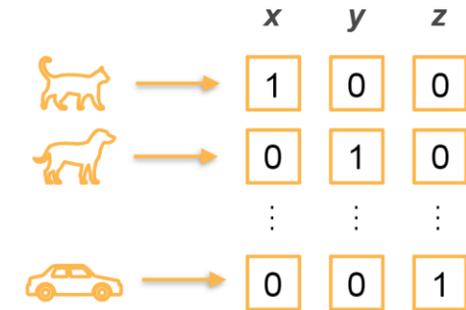
# From Shallow to Deep Language Representations

- GluonNLP (MXNet)
- Dados de Palavras
  - Word2Vec
  - Bag of Words
  - FastText
  - GloVe
- Uso
  - Similaridades
  - Analogias
  - Análise de Sentimentos

## One-Hot Encoding

- One-hot vectors map objects (words) to fixed-length vectors
- Vectors contain only ID, **no semantic meaning**

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{z}, \mathbf{y} \rangle = 0$$

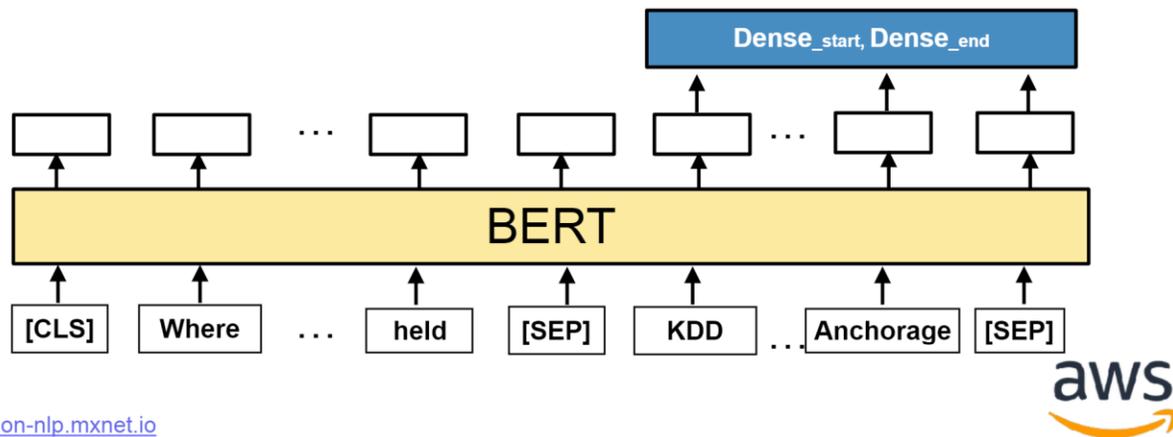


# From Shallow to Deep Language Representations

- BERT (<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>)

## Fine-tuning: Question Answering

Input\_0: KDD 2019 is held in Anchorage  
Input\_1: Where is KDD held  
Output: Anchorage



[gluon-nlp.mxnet.io](http://gluon-nlp.mxnet.io)

## BERT

Bidirectional Embedding from Transformers

[gluon-nlp.mxnet.io](http://gluon-nlp.mxnet.io)

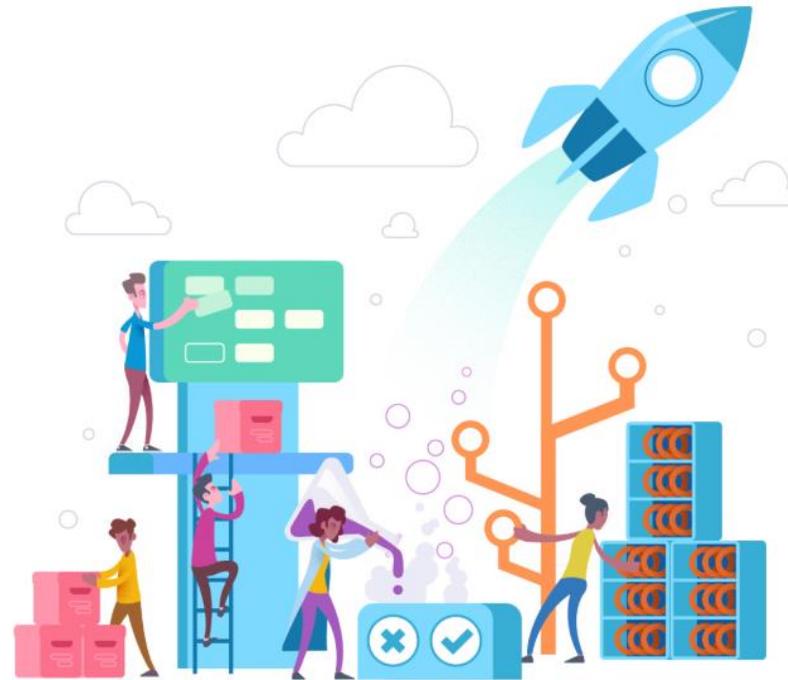


# Democratizing & Accelerating AI through Automated Machine Learning

## Azure Machine Learning studio (Preview)

Quickly prep data, train, and deploy machine learning models.  
Improve productivity and costs with autoscaling compute and pipelines.

[Sign in](#)





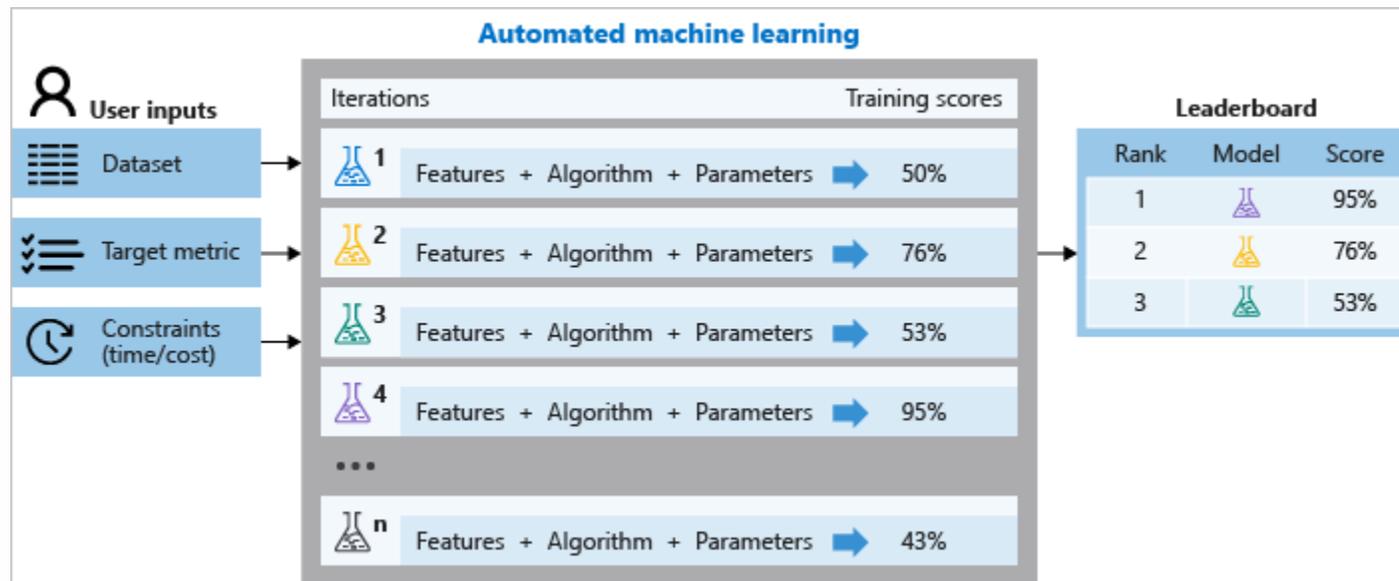
# Democratizing & Accelerating AI through Automated Machine Learning

- Azure AutoML
  - Implementar soluções de Machine Learning sem amplo conhecimento de programação;
  - Economize tempo e recursos;
  - Aproveitar as práticas recomendadas de ciência de dados;
  - Fornecer solução de problemas ágil;

A tabela a seguir lista casos comuns de uso de ML automatizados.

| Classificação         | regressão                            | Previsão de série temporal |
|-----------------------|--------------------------------------|----------------------------|
| Detecção de fraudes   | Previsão de desempenho da CPU        | Previsão de demanda        |
| Previsão de marketing | Previsão de durabilidade de material | Previsão de vendas         |

# Democratizing & Accelerating AI through Automated Machine Learning





# Deep Learning at Scale on Databricks

- Keras and Neural Network Fundamentals
  - MLflow and Spark UDFs
  - Horovod: Distributed Model Training
  - LIME, SHAP & Model Interpretability
- 
- <https://brookewenig.github.io/DeepLearning.html#/1>

# Deep Learning at Scale on Databricks



## Python: The Language of Deep Learning?

```
with tf.variable_scope('conv1') as scope:
    kernel = _variable_with_weight_decay('weights',
                                        shape=[3, 3, 3, 32],
                                        dtype=tf.float32,
                                        initializer=tf.random_normal_initializer(0.01))
    conv1 = tf.nn.conv2d(images, kernel, [1, 1, 1, 1], padding='SAME')
    biases = _variable_on_cpu('biases', [32], tf.constant_initializer(0.01))
    pre_activation = tf.nn.conv2d(images, kernel, [1, 1, 1, 1], padding='SAME')
    conv1 = tf.nn.relu(tf.nn.conv2d(images, kernel, [1, 1, 1, 1], padding='SAME') +
                     tf.nn.bias_add(pre_activation, biases))
    activation_summary(conv1)

# pool1
pool1 = tf.nn.max_pool(conv1, [1, 3, 3, 1], strides=[1, 2, 2, 1],
                        padding='SAME', name='pool1')

# conv2
kernel = tf.nn.conv2d(pool1, 4, [1, 3, 3, 1], strides=[1, 2, 2, 1],
                      padding='SAME')

# conv2
with tf.variable_scope('conv2') as scope:
    kernel = _variable_with_weight_decay('weights',
                                        shape=[3, 3, 64, 32],
                                        dtype=tf.float32,
                                        initializer=tf.random_normal_initializer(0.01))
    conv2 = tf.nn.conv2d(pool1, kernel, [1, 1, 1, 1], padding='SAME')
    biases = _variable_on_cpu('biases', [32], tf.constant_initializer(0.01))
    pre_activation = tf.nn.conv2d(pool1, kernel, [1, 1, 1, 1], padding='SAME')
    conv2 = tf.nn.relu(tf.nn.conv2d(pool1, kernel, [1, 1, 1, 1], padding='SAME') +
                     tf.nn.bias_add(pre_activation, biases))
    activation_summary(conv2)

# conv3
```

TensorFlow

```
model = Sequential()
model.add(Conv2D(32, (3, 3), padding='same',
                input_shape=x_train.shape[1:]))
model.add(Activation('relu'))
model.add(Conv2D(32, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))

model.add(Conv2D(64, (3, 3), padding='same'))
model.add(Activation('relu'))
model.add(Conv2D(64, (3, 3)))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))

model.add(Flatten())
model.add(Dense(512))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(num_classes))
model.add(Activation('softmax'))
```

Keras

```
from torch.autograd import Variable
import torch.nn as nn
import torch.nn.functional as F

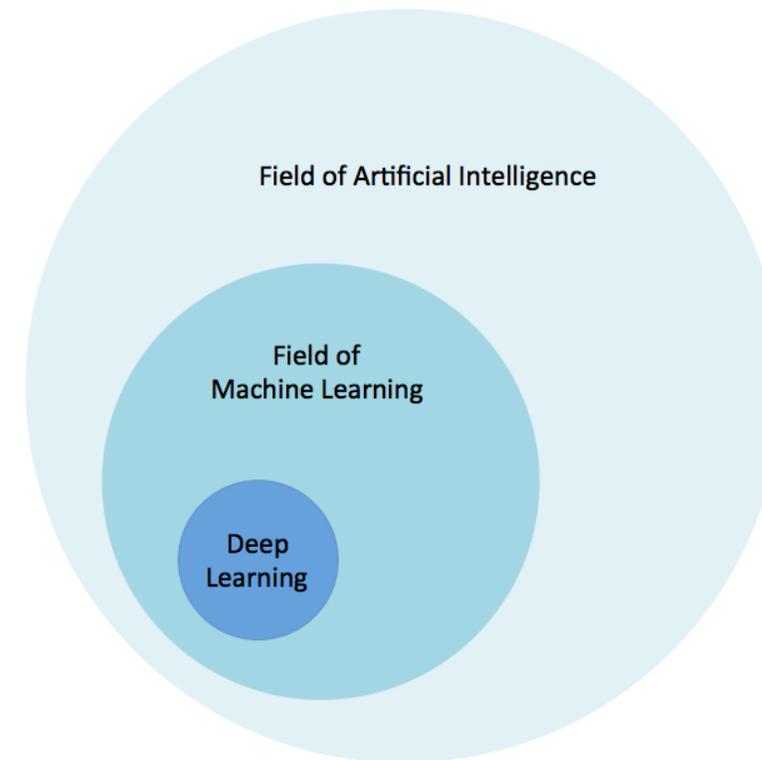
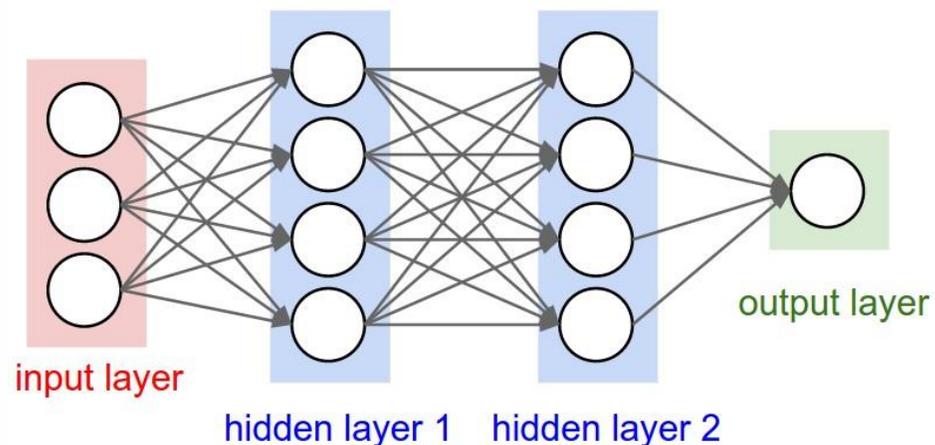
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(3, 5)
        self.pool = nn.MaxPool2d(2, 2)
        self.conv2 = nn.Conv2d(5, 10)
        self.fc1 = nn.Linear(16 * 5 * 5, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)

    def forward(self, x):
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        x = x.view(-1, 16 * 5 * 5)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x
```

PyTorch

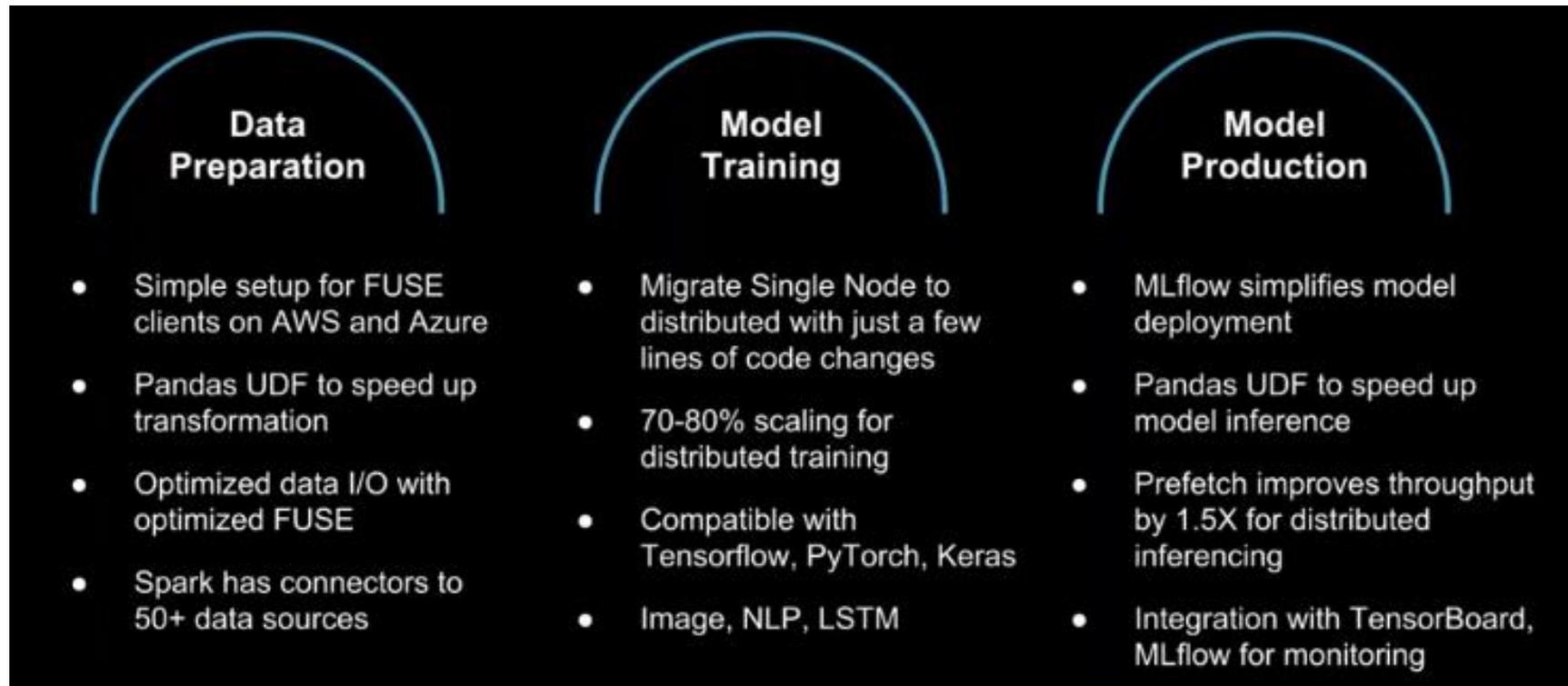
# Deep Learning at Scale on Databricks

- Por que Deep Learning?
  - Análise de imagens
  - Linguagem Natural
- Escalável.
- Aceita vários tipos de dados.





# Deep Learning at Scale on Databricks



# Lições aprendidas

- O processo de solicitação para participação em eventos de curta e média duração da CGU.
- O privilégio de participar de um evento tão importante.
- A importância de se estruturar iniciativas de análise de dados na CGU.
- A necessidade de se capacitar servidores da CGU.



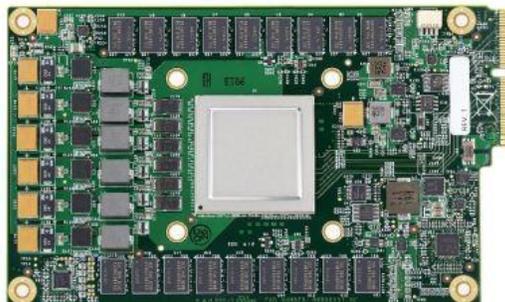
# Oportunidades para a CGU



- Padronizar desenvolvimento de modelos de ML.
  - Rapidez no desenvolvimento.
  - Facilidade de manutenção.
- Discutir a criação de um grupo de dados institucional.
- Usar ferramentas/bibliotecas de automação de ML.
- Aumentar a divulgação do portal webODP:
  - <https://webodp.cgu.local/>.
- Integrar/disponibilizar modelos já existentes (plataforma de modelos).
  - MARA – Mapa de Riscos (<https://webodp/trilhas/mara/>)
  - PNAE - Programa Nacional de Alimentação Escolar (<https://webodp.cgu.local/trilhas/pnae/>)
  - ALICE
- Uso de componentes dedicados pra ML:
  - GPUs (CGU já tem);
  - Nuvem (problema no sigilo dos dados);
- Ampliar o uso do Cluster Hadoop da CGU.
  - Espaço físico compartilhado total de 40TB.
  - Espaço físico pessoal de 1,5TB.
  - 5 máquinas:
    - 1 TB de RAM cada.
    - 56 núcleos.

- TPU Card to replace a disk
- Up to 4 cards / server

TPU Card & Package



Obrigado!



Rodrigo Dewes (SCC/DIE)

[rodrigo.dewes@cgu.gov.br](mailto:rodrigo.dewes@cgu.gov.br)

Rodrigo Pedatella (SFC/DAE)

[rodrigo.pedatella@cgu.gov.br](mailto:rodrigo.pedatella@cgu.gov.br)