CGU Controladoria-Geral da União

Ciência de Dados

Estado da Arte e Oportunidades para a CGU

Rodrigo Dewes (SCC/DIE)
rodrigo.dewes@cgu.gov.br
Rodrigo Pedatella (SFC/DAE)
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

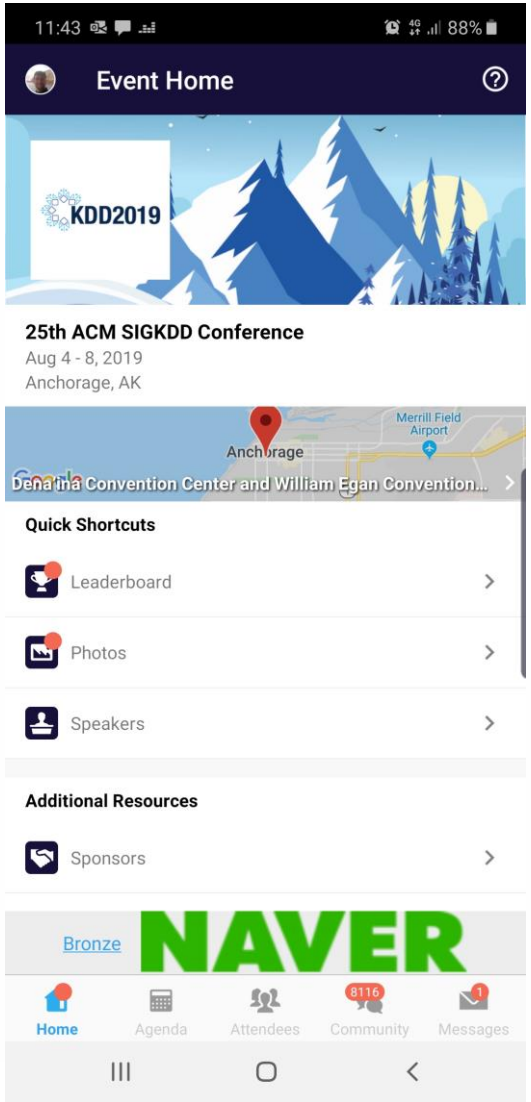
O que é o KDD?

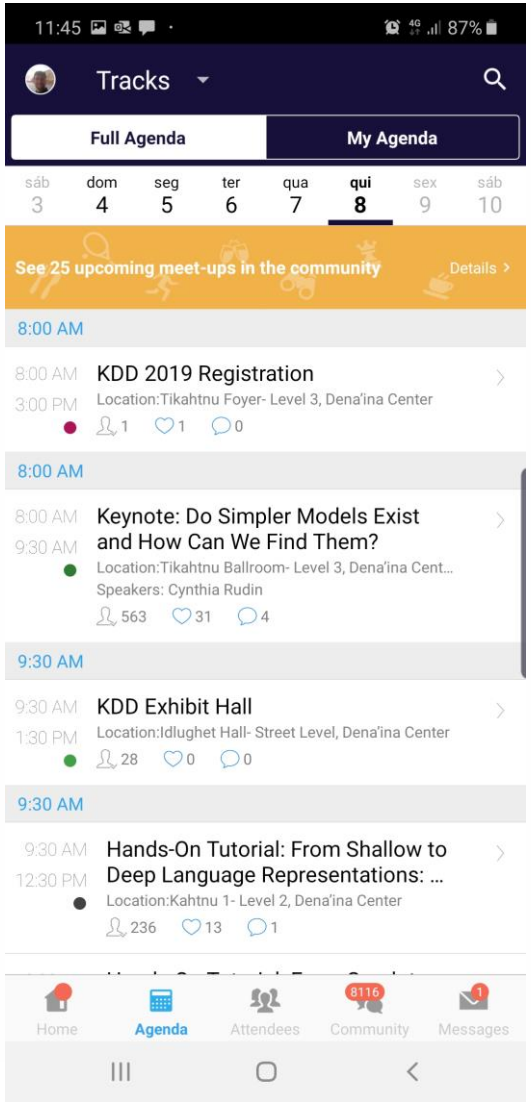
- Mais importante evento sobre ciência de dados do mundo.
- Ocorre anualmente, sob organização da ACM.
- Reúne os maiores pesquisadores do **meio acadêmico**, **indústria** e **governo** especializados em *data science*, *data mining*, *knowlegde discovery*, *large-scale data analytics* e *big data*.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

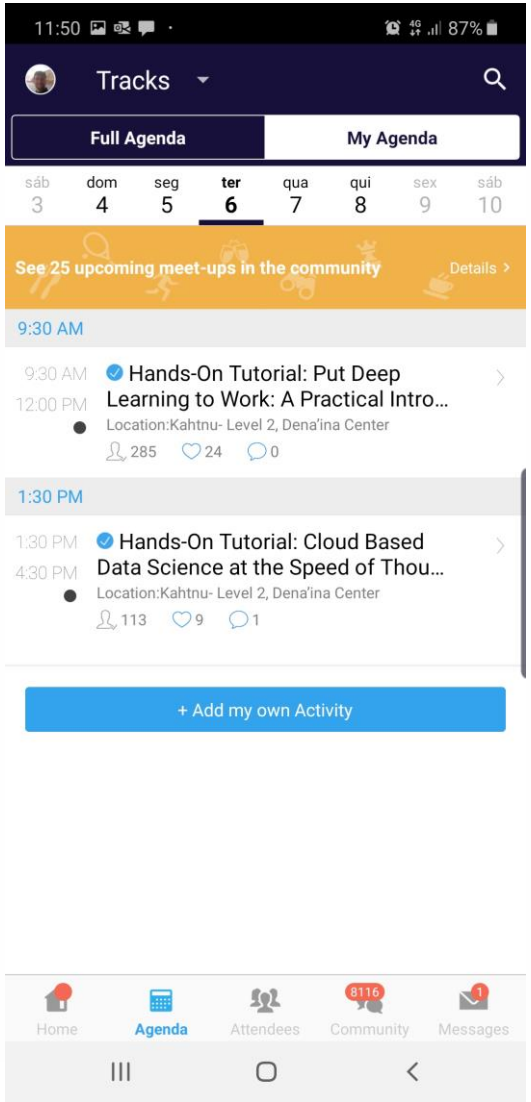
KDD 2019 em números

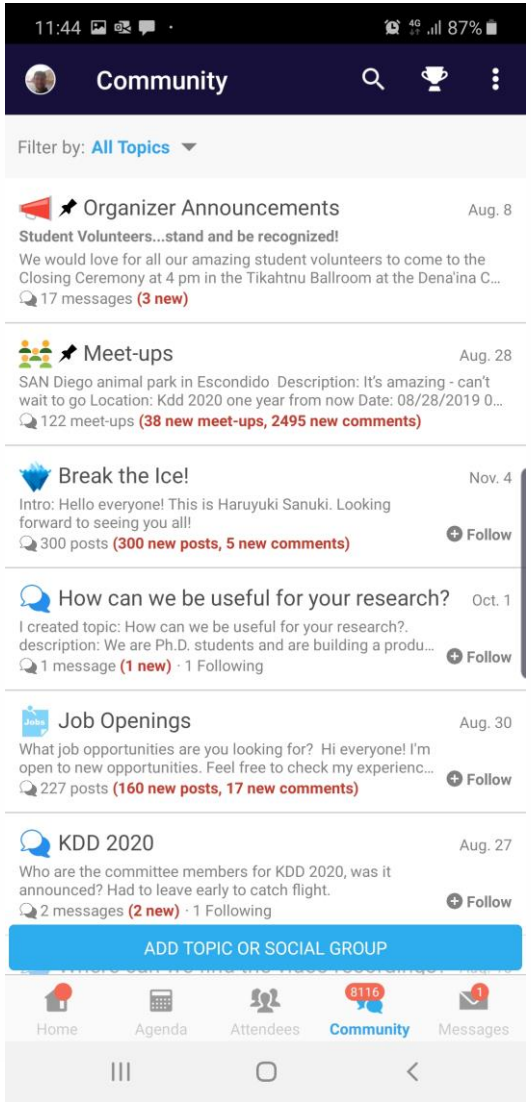
- 5 dias de atividades (04/08 a 08/08).
- 3 mil participantes.
- 2 centros de convenções:
  - Dena'ina Convention Center.
  - William Egan Convention Center.
- 2 keynote speakers e 1 plenary keynote panel.
  - Mais importantes cientistas de dados do mundo compartilham seu conhecimento para o avanço da 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









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 exibições 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*

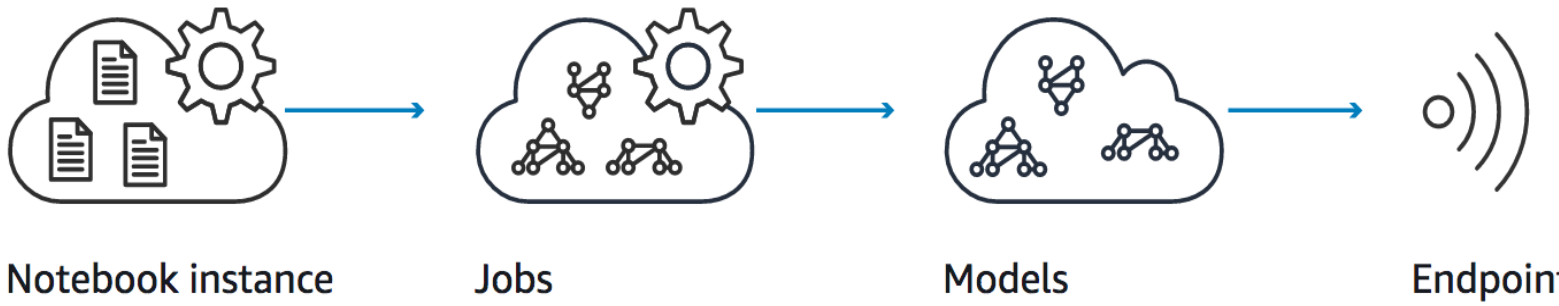
Participamos dos seguintes tutoriais:

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

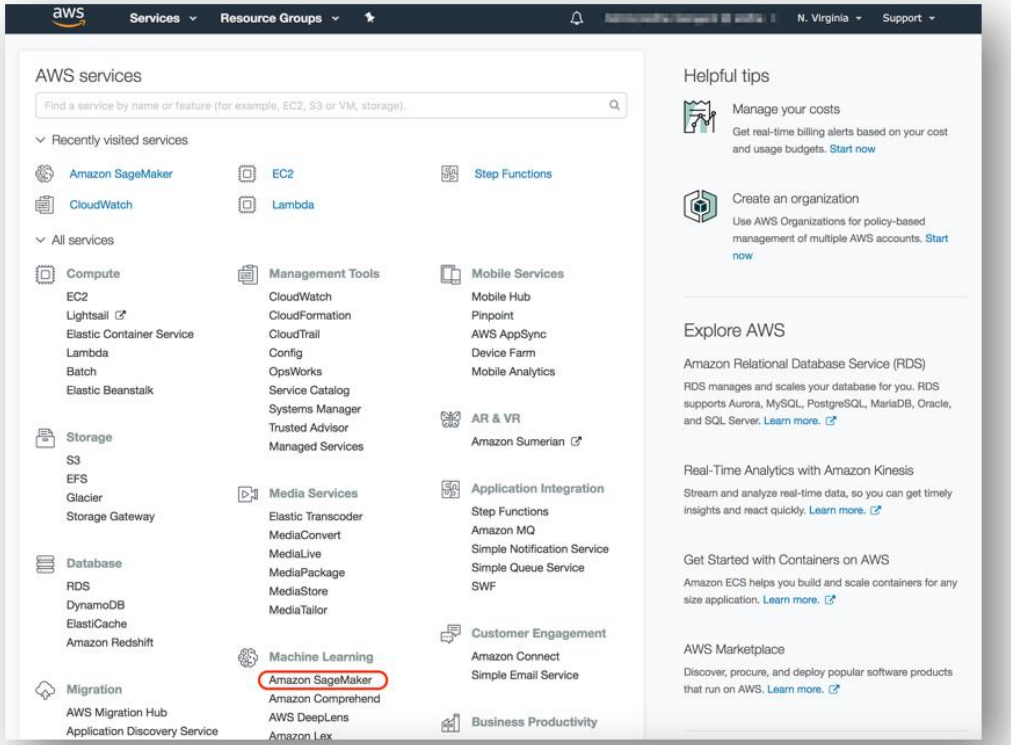
https://www.kdd.org/kdd2019/hands-on-tutorials

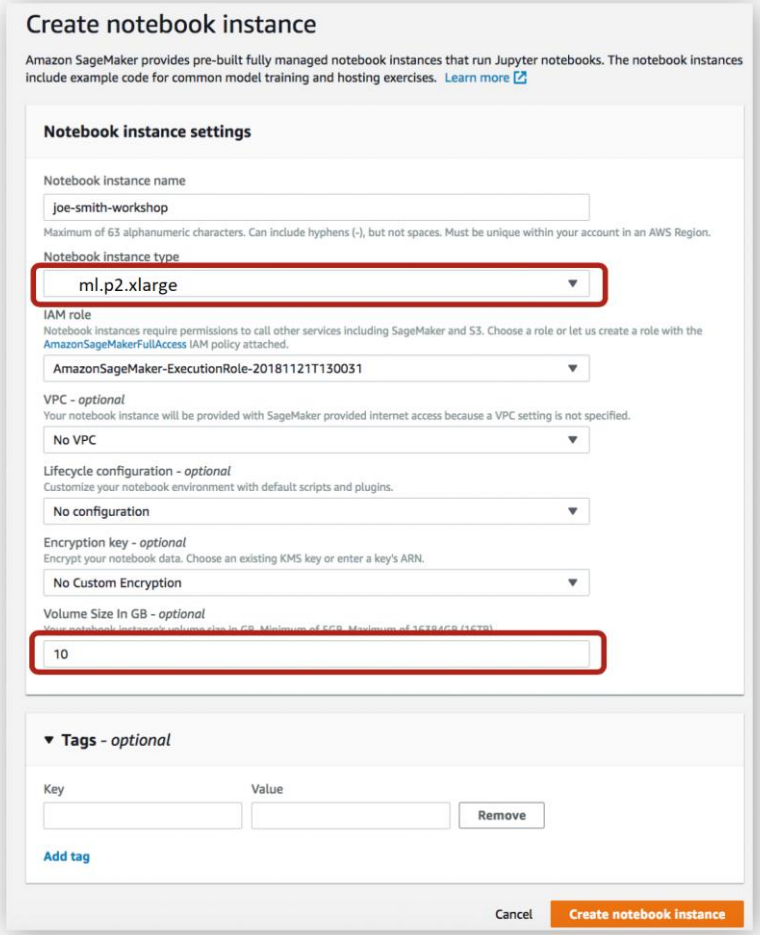
KDD2019 Amazon SageMaker Labs

- 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





KDD2019 Amazon SageMaker Labs

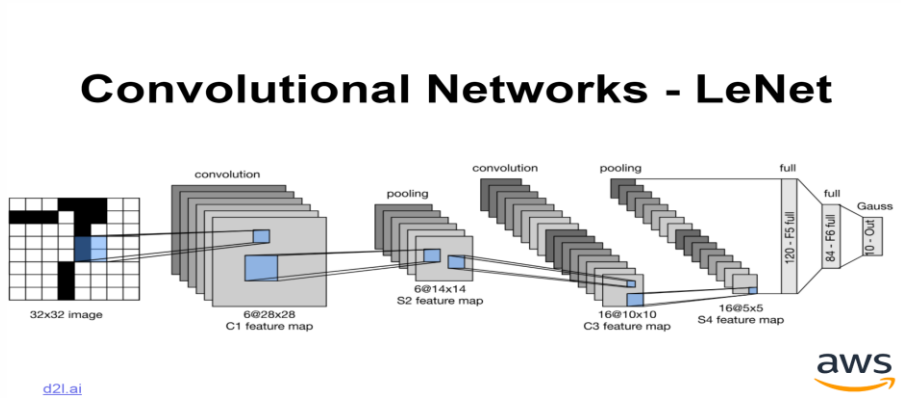
- Classificou a imagem abaixo como pessoa.

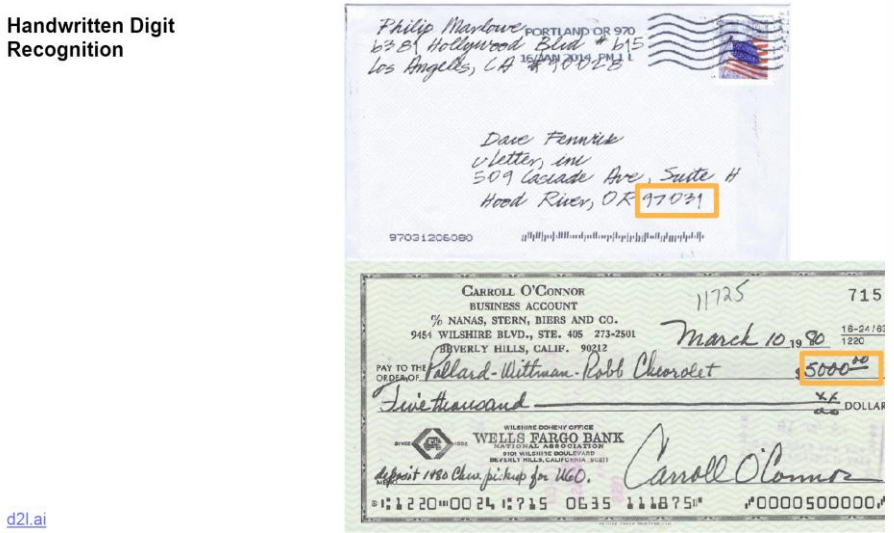
NVidia RAPIDS

- 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

From Shallow to Deep Language Representations

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





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**

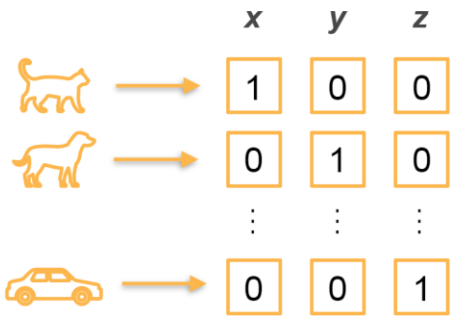
From Shallow to Deep Language Representations

- 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**

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

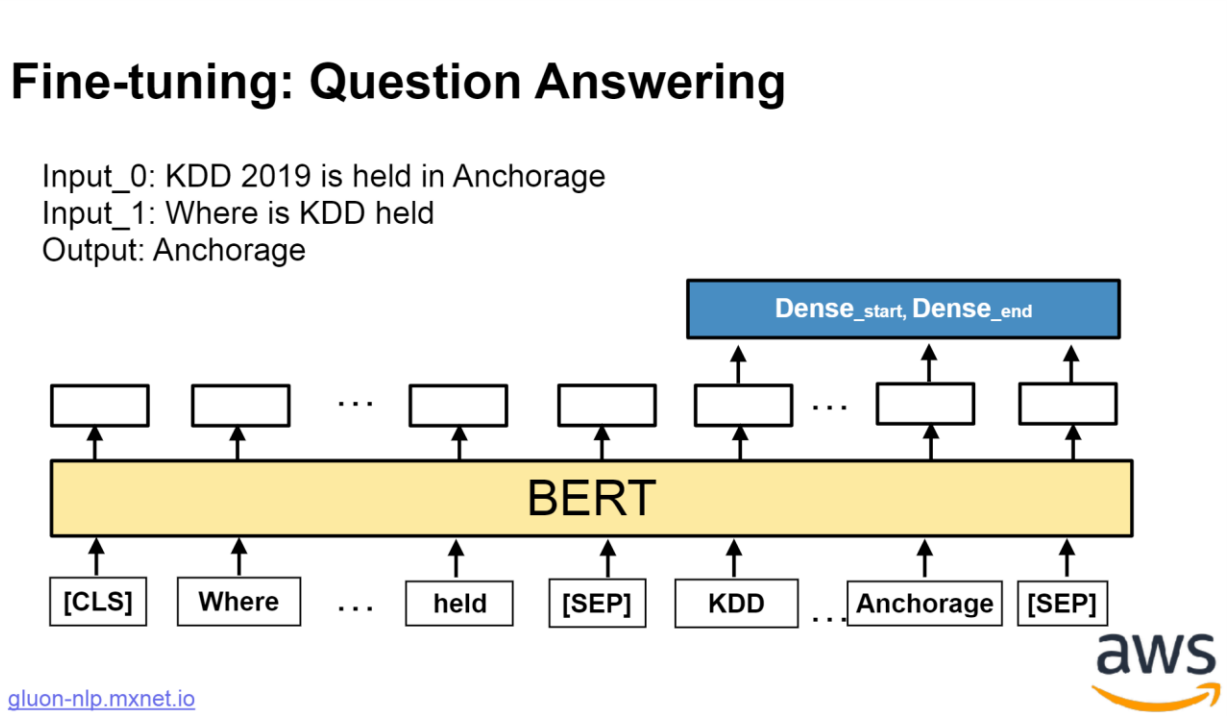


d2l.ai

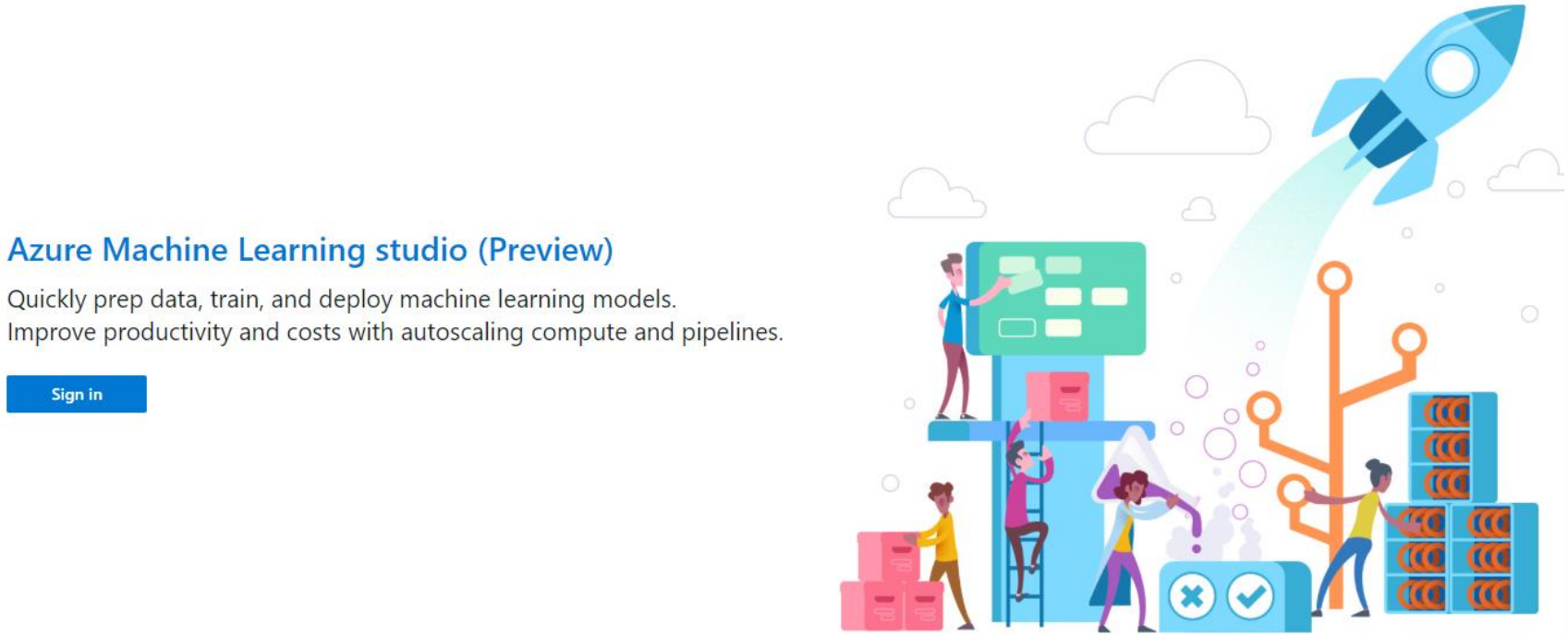
aws

From Shallow to Deep Language Representations

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



Democratizing & Accelerating AI through Automated Machine Learning



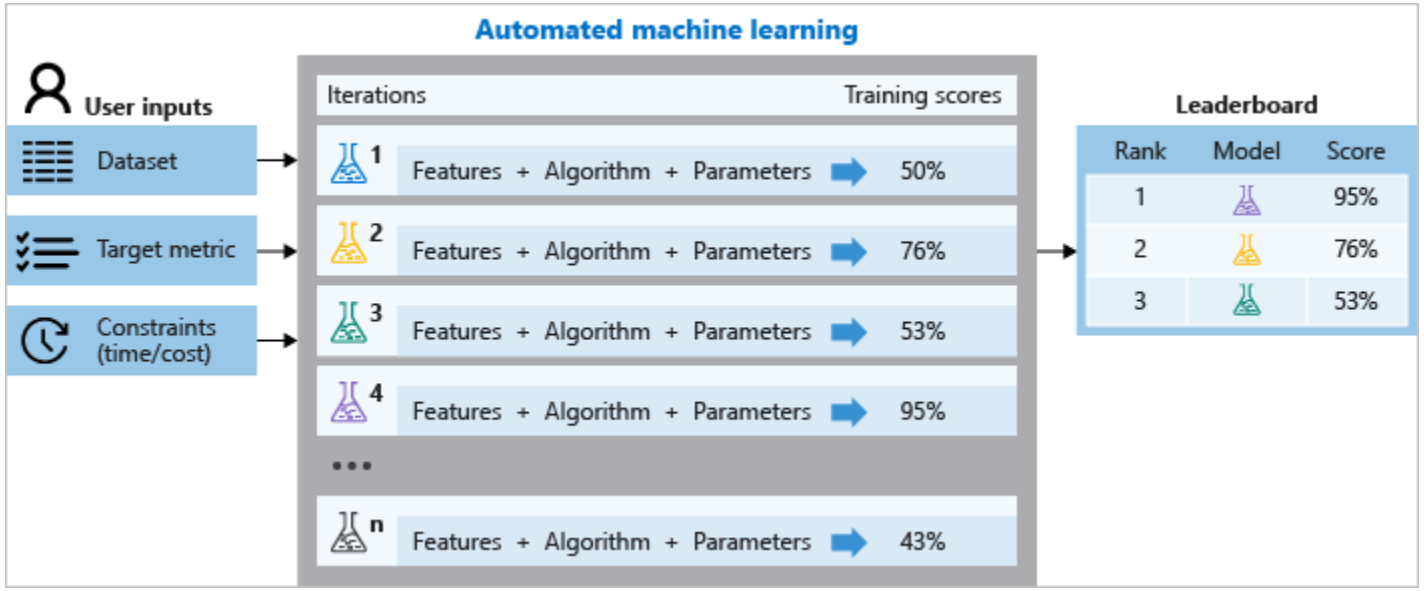
Democratizing & Accelerating AI through Automated Machine Learning

- 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

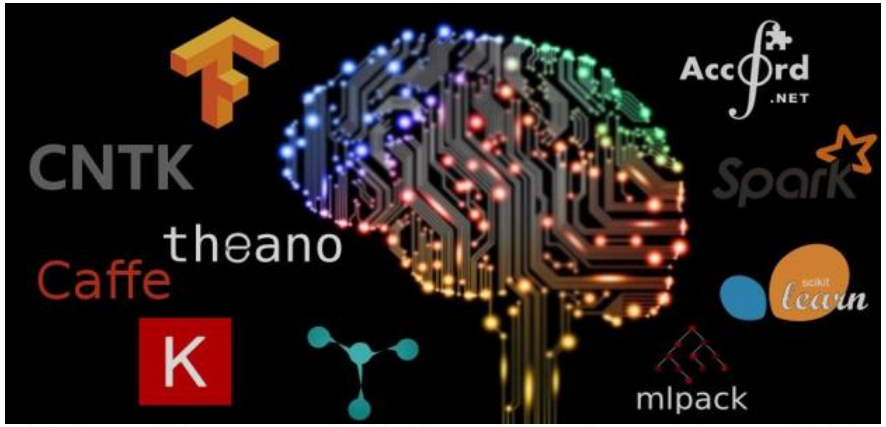


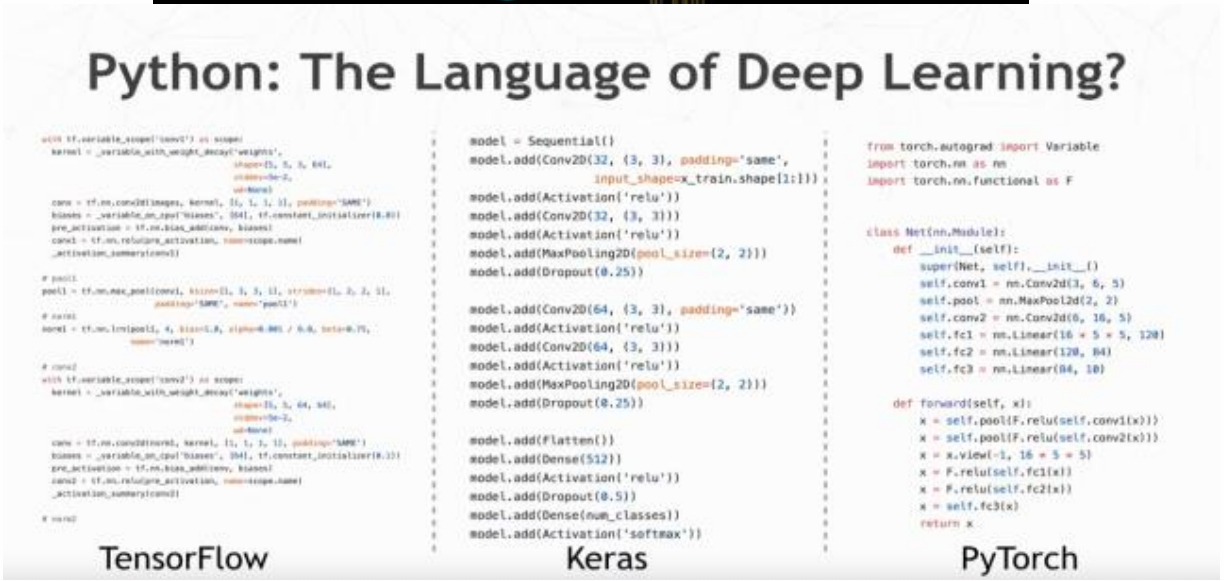
Deep Learning at Scale on Databricks

- Keras and Neural Network Fundamentals
- MLflow and Spark UDFs
- Horovod: Distributed Model Training
- LIME, SHAP & Model Interpretability

- https://brookewenig.github.io/DeepLearning.html#/1

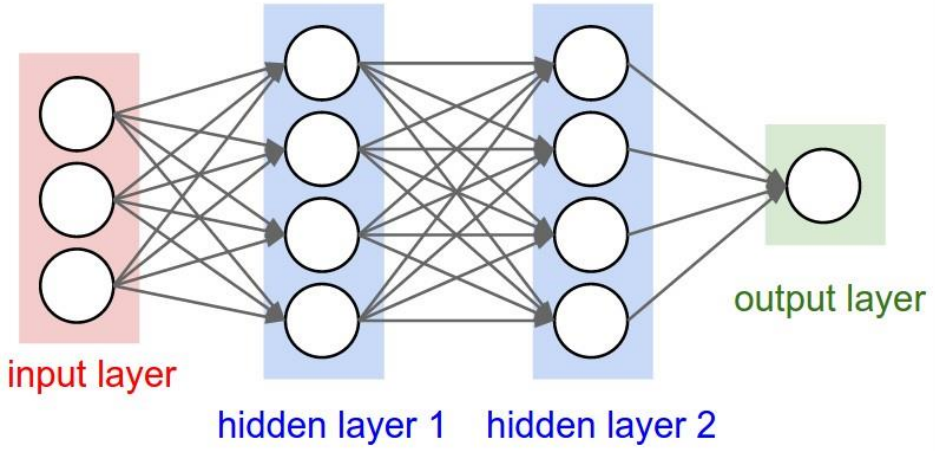
Deep Learning at Scale on Databricks

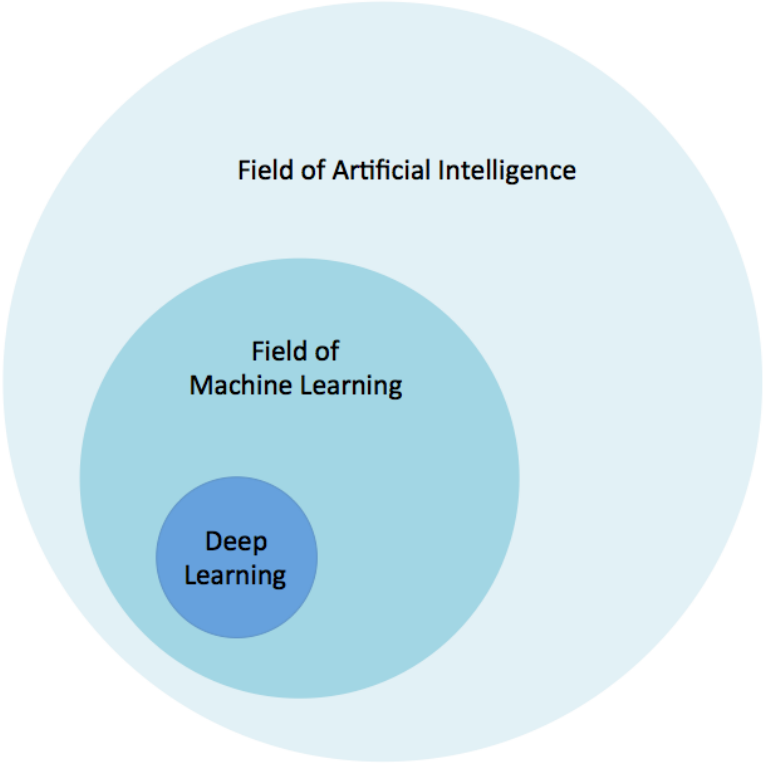




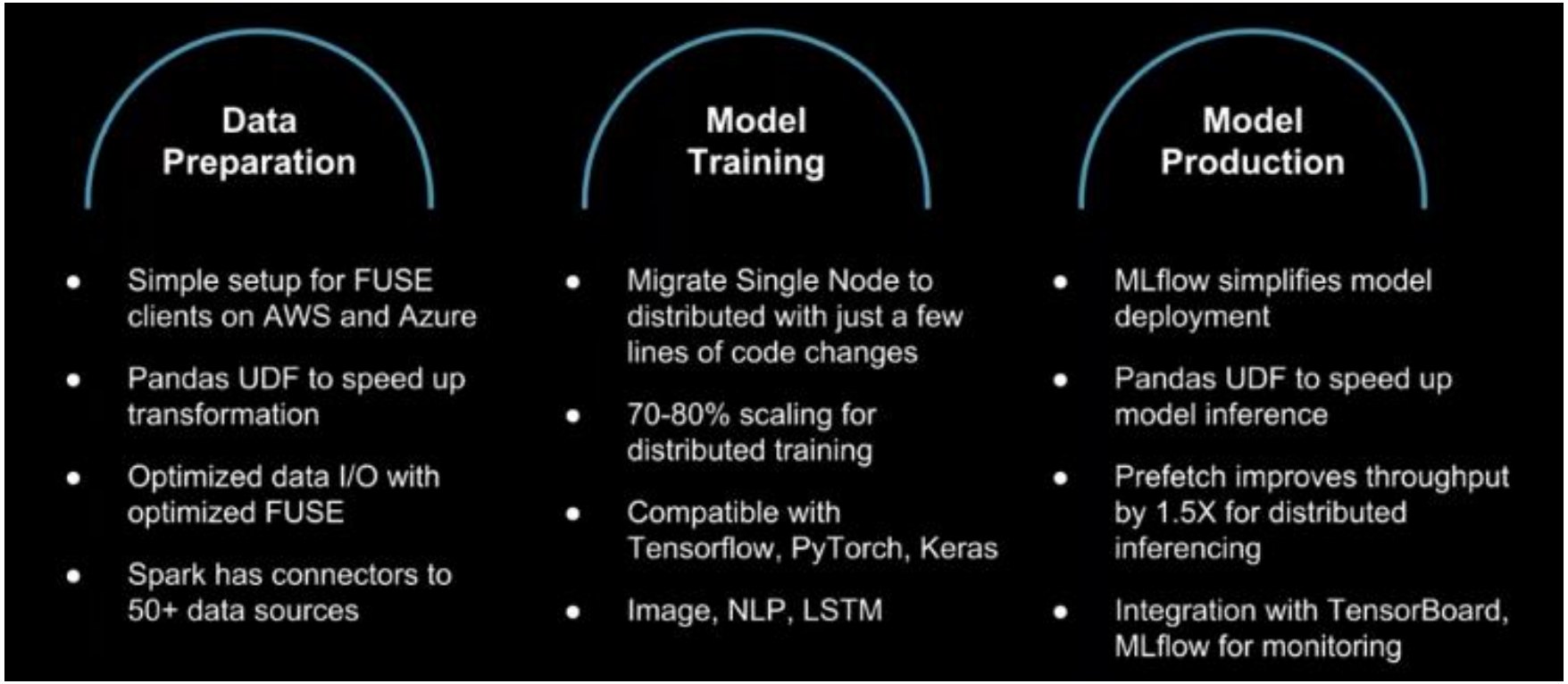
Deep Learning at Scale on Databricks

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





Deep Learning at Scale on Databricks



Lições aprendidas

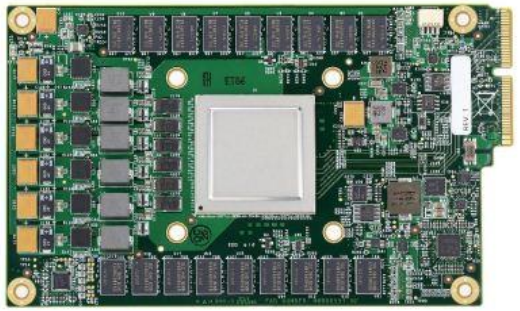
- O processo de solicitação para participação em eventos de curta e média duração da CGU.
- O privilégio de participar de um evento tão importante.
- A importância de se estruturar iniciativas de análise de dados na CGU.
- A necessidade de se capacitar servidores da CGU.

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
Rodrigo Pedatella (SFC/DAE)
rodrigo.pedatella@cgu.gov.br