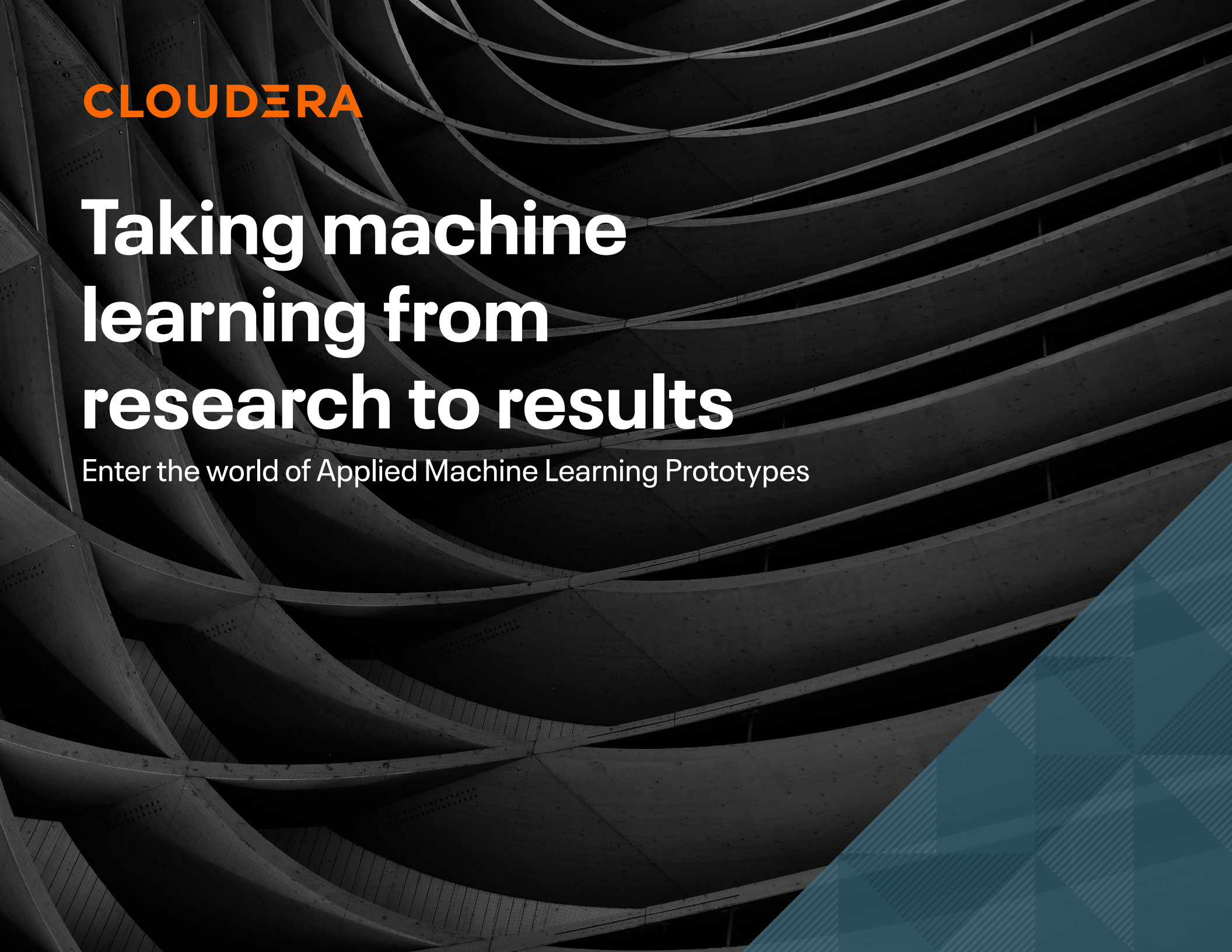

CLOUDERA
Taking machine learning from research to results
Enter the world of Applied Machine Learning Prototypes

Table of Contents

Introduction: What are AMPs?

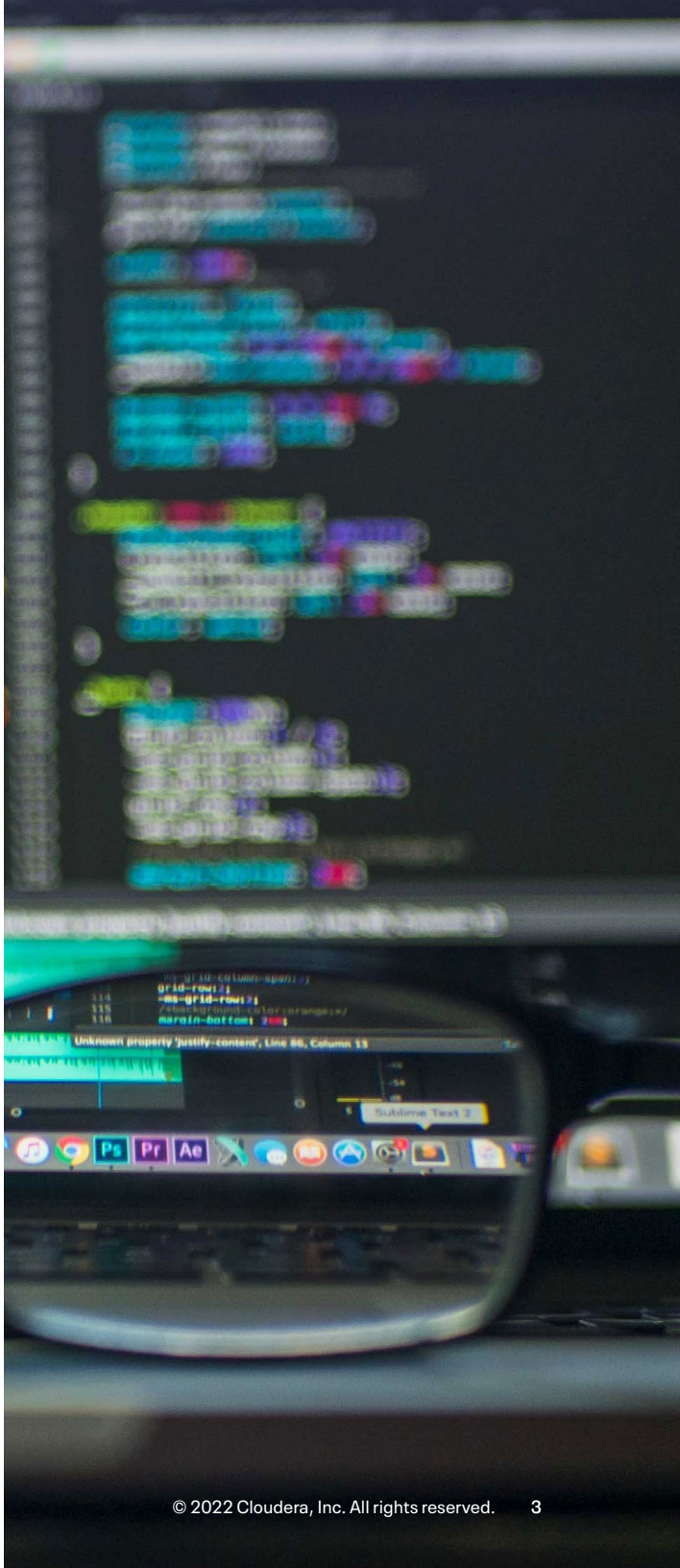
The field of data science has benefited greatly from open source collaboration. Progress has accelerated thanks to the open availability of useful code and the spirit of community that encourages sharing knowledge.

In the world of data science today, no one starts with a blank script, and a big benefit of open source to data practitioners is access to tutorials and example projects. Though the reality is that not all code is the same quality. Using code from a random tutorial doesn't mean that it is based on current methodologies, is credible, or even works.

To support open source machine learning (ML) practitioners, Cloudera released a library of Applied Machine Learning Prototypes (AMPs), fully developed solutions that offer data scientists a running start to bring ML applications from concept to reality. **Concepts are proven. The code is tested. Dependencies are documented and available.** By learning from prototypes and modifying them, data scientists have achieved results faster.

AMPs support rapid delivery, great reliability, and full customization, all adding up to a greater likelihood of project success. Cloudera Machine Learning (CML) users on Cloudera Data Platform (CDP) can deploy AMPs with one click to tackle common use cases. True to the open source community spirit, AMPs are available to anyone.

They are created by Cloudera Fast Forward Labs (CFFL), the research group within Cloudera that follows emerging trends in ML and develops free research reports for the data science community. AMPs are open source yet aren't accompanied by the risk that can sometimes be associated with open source. With AMPs, you'll get the same thing every time—high-quality, professionally developed code.



AMPs highlights:

FAST

AMPs aren't tutorials or demos; they are working ML use cases that launch with a single click. AMPs lower the barrier to entry for ML development teams. They give data scientists a starting point to build, deploy, and monitor business-ready ML applications. Delivering ML models becomes a fast effort with access to the AMP catalog.

RELIABLE

The researchers who create AMPs work at the leading edge of ML innovation. Best practices are built into every AMP, improving security and removing guesswork. AMPs go through a rigorous testing and review process before release.

CUSTOMIZABLE

Each AMP offers a full end-to-end ML framework for a common use case. Data scientists can retrain a model on different datasets, building ML applications unique to their organization and industry. Parts of AMPs can be applied to other projects. This versatility makes it easier for organizations to use ML to a competitive advantage.

In an AMP, you may find:

- **Python scripts**
  Code that performs actions in support of the ML use case
- **Jupyter notebooks**
  Hands-on code examples with explanations
- **Sample data sets**
  Data to get started
- **Project Specification File**
  Instructions to facilitate automated project setup
- **Web application**
- **Deployed model**
- **Documentation**
  Including links to helpful resources

AMP: Deep Learning for Image Analysis

Machine learning and images: The perfect match

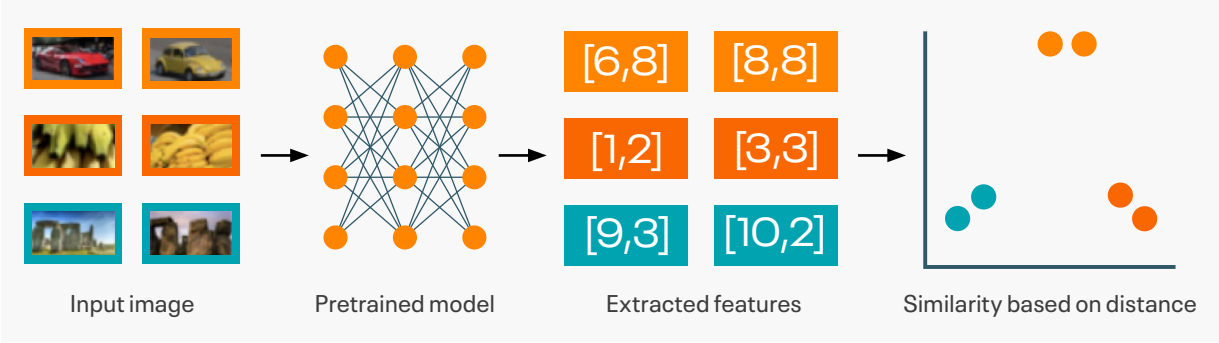
Image analysis is one of the great **success stories** of deep learning. It's intuitive and supported by real-life examples of ML apps that identify anomalies in medical images, spot defects in products on assembly lines, or classify images in large collections.

With this success comes risk: It may seem easy, but that impression of ease can create blind spots for how this tech can fail. Ethical concerns around training data bias, privacy violations, and the potential to generate realistic fake images also pose potential pitfalls.

PyTorch and TensorFlow are among the best-known deep learning frameworks successfully applied to image analysis. But with an abundance of choices, it's a daunting challenge to figure out which combination of technologies works best for a given use case.

Building an image analysis AMP

Cloudera researchers have been investigating ways to use deep learning for image analysis since their first report on the technology in 2015. While building the AMP catalog, researchers saw an opportunity to translate this research into a prototype for data scientists to learn from and use when building working, actionable ML models.



Source: Github

CFFL took an existing prototype and made it more easily reproduced by refactoring it to include YAML instructions. The result is the Semantic Image Search with Convolutional Neural Networks AMP. This AMP demonstrates how to **build a scalable semantic search solution on a dataset of images**—finding matches in large image sets. It relies on convolutional neural networks (CNNs) to turn images into semantically meaningful representations, which are then indexed using the FAISS search matching algorithm. The project includes an interactive visualization for exploring the results of searches made using different model architectures.

See the AMP on Github

Read the Fast Forward Labs report

View a live demo of the original prototype

How convolutional neural network models (e.g. EfficientNets) can be applied to the task of semantic search.

Source: Github

Practical use cases:

Product recommendations

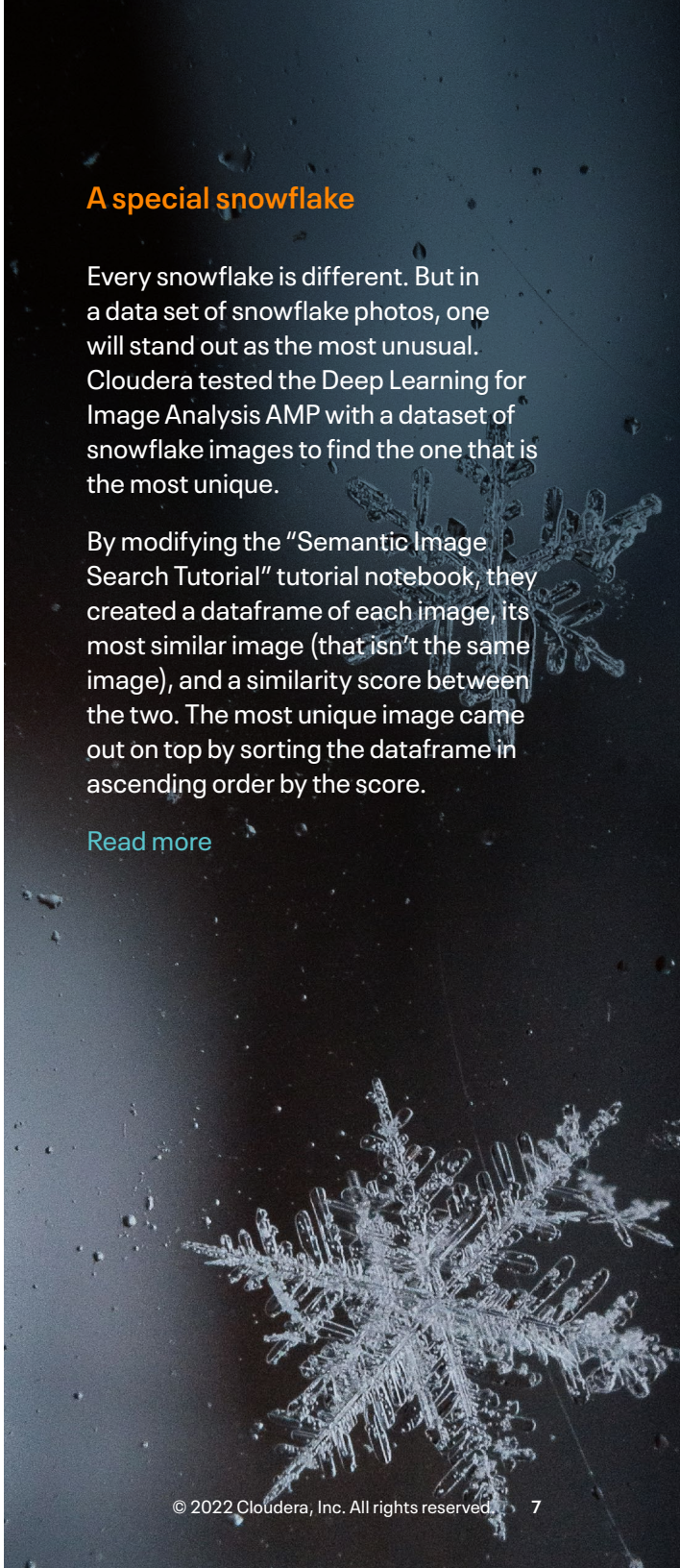
Someone who likes a particular clothing style can be shown pictures of similarly styled items automatically.

Spotting defective products

Manufacturers can use cameras on a factory line to photograph each product, comparing it to an image of a correctly made product, and look for differences. This use case is especially helpful for intricate items like circuit boards since a computer can spot differences that are difficult for humans to see.

Comparing signatures

Compare a new signature image on checks or other documents to an existing record; if the images are similar enough, it's likely authentic.



A special snowflake

Every snowflake is different. But in a data set of snowflake photos, one will stand out as the most unusual. Cloudera tested the Deep Learning for Image Analysis AMP with a dataset of snowflake images to find the one that is the most unique.

By modifying the "Semantic Image Search Tutorial" tutorial notebook, they created a dataframe of each image, its most similar image (that isn't the same image), and a similarity score between the two. The most unique image came out on top by sorting the dataframe in ascending order by the score.

Read more

AMP: Few-Shot Text Classification

Need labels? No problem!

Machine learning excels at classifying text data into categories—typically starting with a large set of labeled data that can be used to train the model.

What happens when you don't have clearly labeled training data? It's possible to use ML techniques to classify documents into categories even if you don't have a set of accurately labeled data. Data scientists call this problem **zero-shot learning,** describing a scenario when a model is trained on one set of labels but then has to sort items into categories it has never seen before. A related concept is **few-shot learning,** in which the model has only a small amount of labeled data for all the labels the model is expected to recognize.

CFFL explored these techniques and found ML methods are effective at categorizing text even without much training data. This research led to the development of the Few-Shot Text Classification AMP.

See the AMP on Github

Read the Fast Forward Labs report

Using an AMP to Classify Text

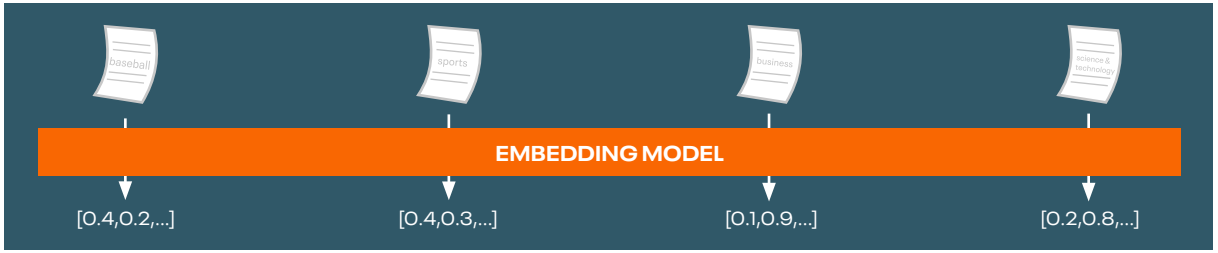
The Few-Shot Text Classification AMP provides a sample user interface that demonstrates how to perform text classification when only a few labeled training examples exist or even when there are no training examples available.

The approach relies on embedding text using word embeddings and sentence embeddings with state-of-the-art Transformer models. No metadata or formatting is required, and this prototype relies on unstructured txt files.

The Few-Shot Text Classification relies on PyTorch and two embedding models: word2vec for words and Sentence-BERT (SBERT) for sentences. It also includes a UI, and you can supply labels for the model to apply to your text data.

Researchers first tested the prototype on two data sets—a set of news articles and Reddit posts—to see how it categorized the text. For example, could the model accurately label all the articles about sports? Interestingly, tests performed best when the models were trained on the most common 20,000 words. Additional words caused classification accuracy to decrease—probably due to rarer words being interpreted as noise.

It may seem counterintuitive, but **sometimes ML can do more with less.** Simple methods can generate good-enough results faster, using fewer computing resources. A complicated model may deliver more accurate results but also take more resources, work much slower, and not be any better for your particular use case.



Practical use cases:

Professional services

Text classification could assist in legal research using a law office's archives. It could also help any company organize documents by department and assist healthcare providers in categorizing documents by health issues or treatments.

Social media content

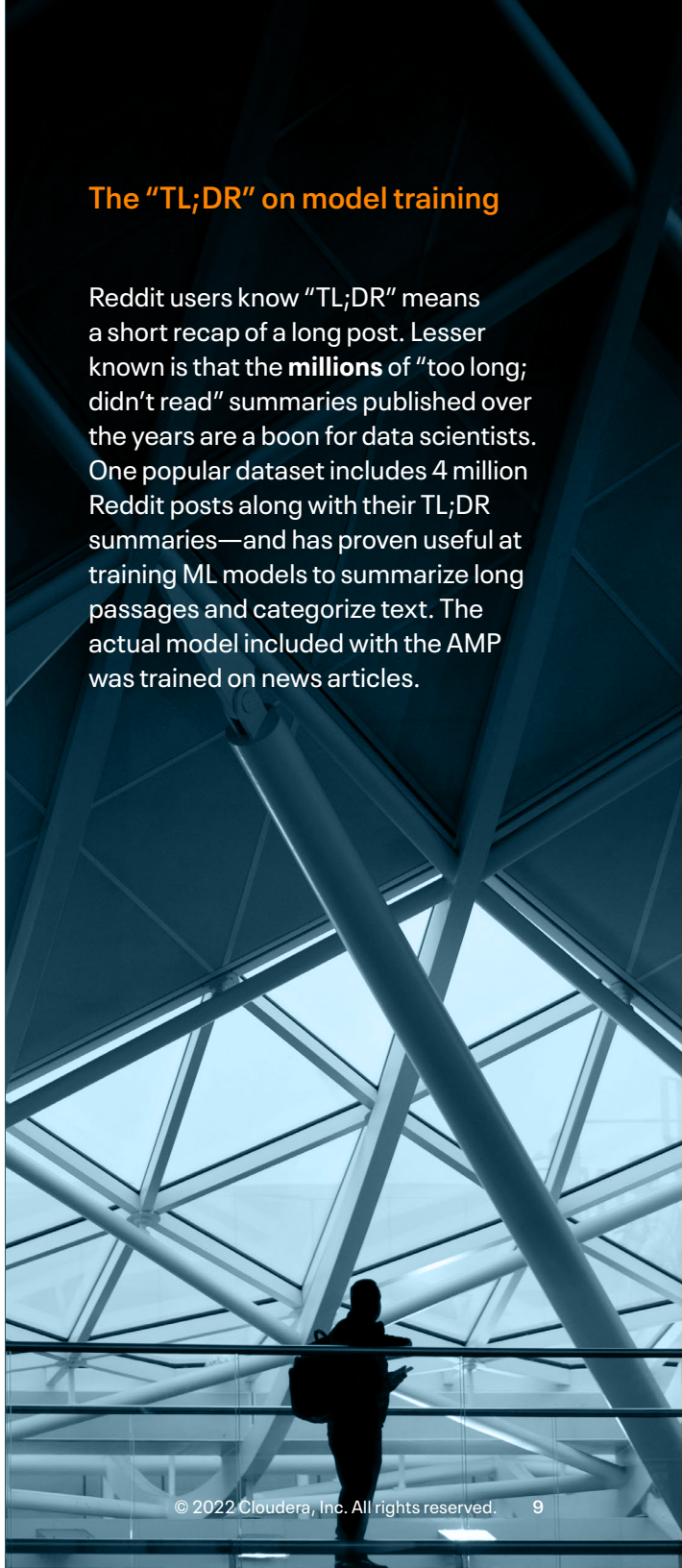
Another potential real-world application is social media, where text content is created so rapidly that humans can't properly categorize it, such as social media postings during a major news or sports event. A social media provider could apply Few-Shot Text Classification as a fast and effective filter to quickly surface messages likely related to the ongoing event.

Recommendation systems

Categorizing products on an e-commerce site, content on a streaming platform is often a first step in order to provide users with recommendations for a type of product or category that the user is interested in. The basis for these classifications can often come from text, such as product descriptions.

The "TL;DR" on model training

Reddit users know "TL;DR" means a short recap of a long post. Lesser known is that the **millions** of "too long; didn't read" summaries published over the years are a boon for data scientists. One popular dataset includes 4 million Reddit posts along with their TL;DR summaries—and has proven useful at training ML models to summarize long passages and categorize text. The actual model included with the AMP was trained on news articles.



Conclusion: Keep Learning

Data science innovates so quickly that it's hard to keep up without help. Having trusted and reliable research from the open source community, and access to working examples like AMPs, is all but essential.

Cloudera AMPs and work from CFFL are helping accelerate what's possible with ML, giving you a jumpstart to get real-world results faster.

Whether you're a Cloudera Machine Learning user or just eager to explore, you can access the full catalog of open source AMPs.

Read more about Cloudera Machine Learning and explore AMPs and other ML use cases.

About Cloudera Fast Forward Labs

Cloudera Fast Forward Labs (CFFL) is an applied machine learning research group. Our mission is to empower enterprise data science practitioners to apply emergent academic research to production machine learning use cases in practical and socially responsible ways, while also driving innovation through the Cloudera ecosystem. Our team brings thoughtful, creative, and diverse perspectives to deeply researched work. In this way, we strive to help organizations make the most of their ML investment as well as educate and inspire the broader machine learning and data science community.

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

CLOUDERA