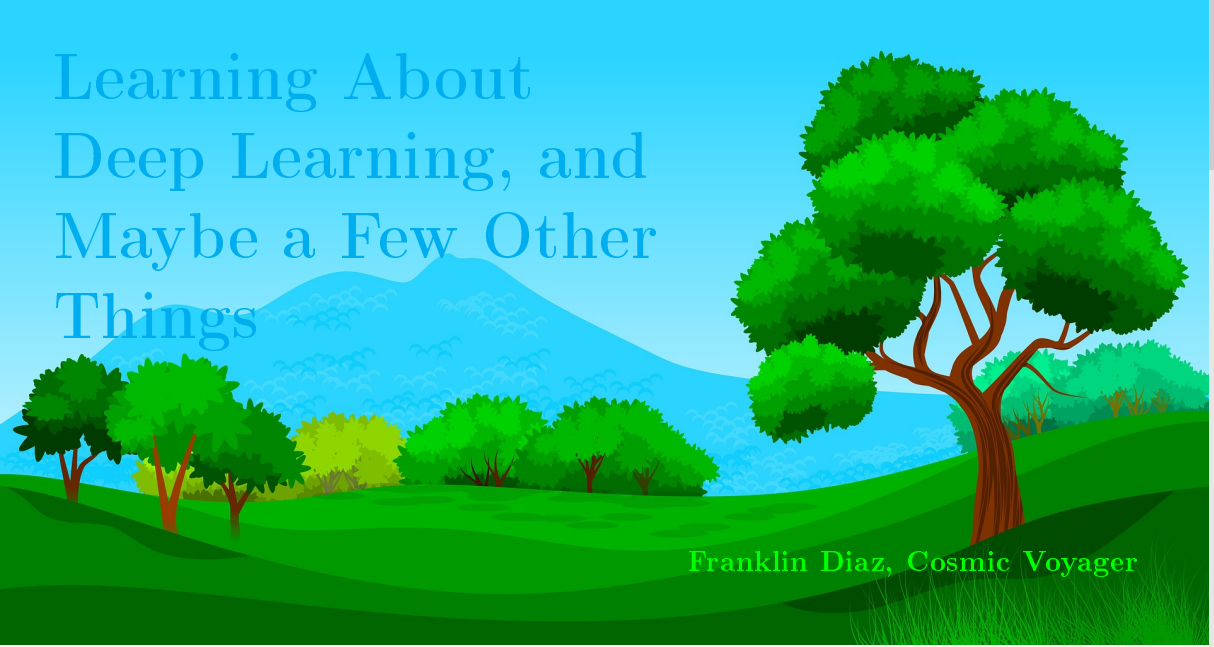
Learning About Deep Learning, and Maybe a Few Other Things



Franklin Diaz, Cosmic Voyager

2022-05-01

- My presentation title is here.
- This talk is about my on-going journey into the world of graph theory and neural networks.
- This is a description of the learning process as well as the project and presenting some results.

•

•

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

This first section is an Introduction to me and my project

About Me

- I am a Security Consultant at Palo Alto Networks, cloud and automation for past 2+ years
- Did Data Eng/DevSecOps at Salesforce for 5 years.
- Been going to security conferences for a while.

- I am a Security Consultant at Palo Alto Networks, cloud and automation for past 2+ years
- Did Data Eng/DevSecOps at Salesforce for 5 years.
- Been going to security conferences for a while.

- Here is a picture of me, modified by a popular local artist.
- In my current role as a consultant, I get to work with the major cloud providers.
- In the past I was not a Data Scientist, but did some time on the Security Data Engineering team at Salesforce. This gave me a bit of a head start with data pipelines, directed acyclic graphs, and a few other things.

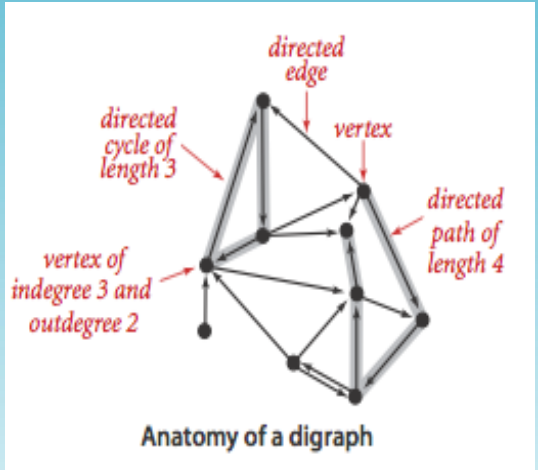
The Project

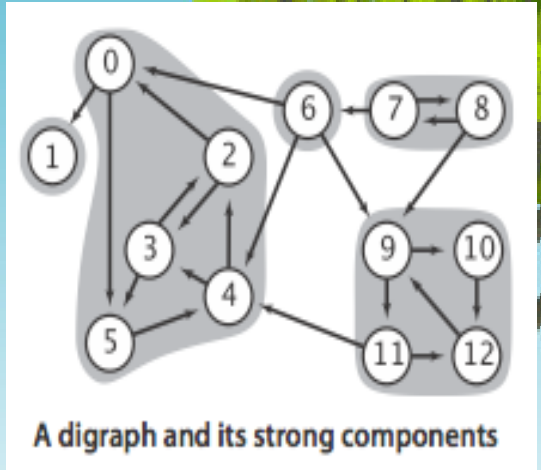
- Realized that Terraform can output directed graphs.
- Had done a lot of work at Salesforce with directed graphs, data pipeline orchestration with AirFlow, etc. so I was somewhat familiar with the output I was seeing.
- The first question I had was, what can I do with these directed graphs?
- My hunch was I could "do some processing and analysis" of all this security infrastructure graph data and hoped that could lead to... predictions?
- All of the Code for (almost) everything is Here

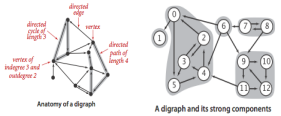
- Realized that Terraform can output directed graphs.
- Had done a lot of work at Salesforce with directed graphs, data pipeline orchestration with AirFlow, etc. so I was somewhat familiar with the output I was seeing.
- The first question I had was, what can I do with these directed graphs?
- My hunch was I could "do some processing and analysis" of all this security infrastructure graph data and hoped that could lead to... predictions?
- All of the Code for (almost) everything is Here

- In case you are not familiar, Terraform is software that allows you to declare resources like network elements in public cloud providers.
- Had and still have this vague notion that if I had enough data I could find "outliers". Maybe like a modernized version of a Pareto analysis?

What's a DiGraph?







- The big takeaway here is the idea of "edges" and "nodes"
- Source: Algorithms, 4th Edition, by Robert Sedgewick and Kevin Wayne
- Wrath of Math!

•

•

Introduction

What is Deep Learning, Exactly?

The Journey

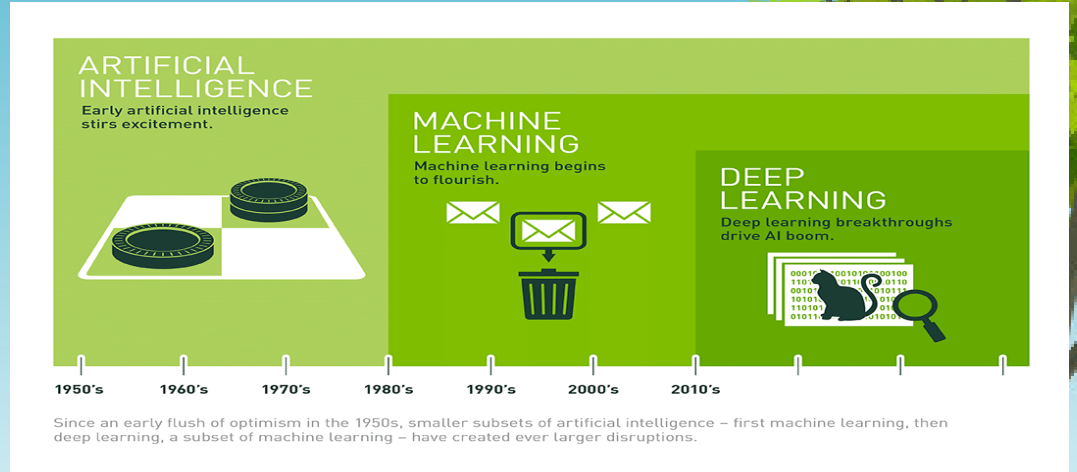
Visualizations

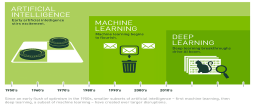
So Now What?

Sources and Citations

- You've probably heard this term lately, lets talk about what it means.

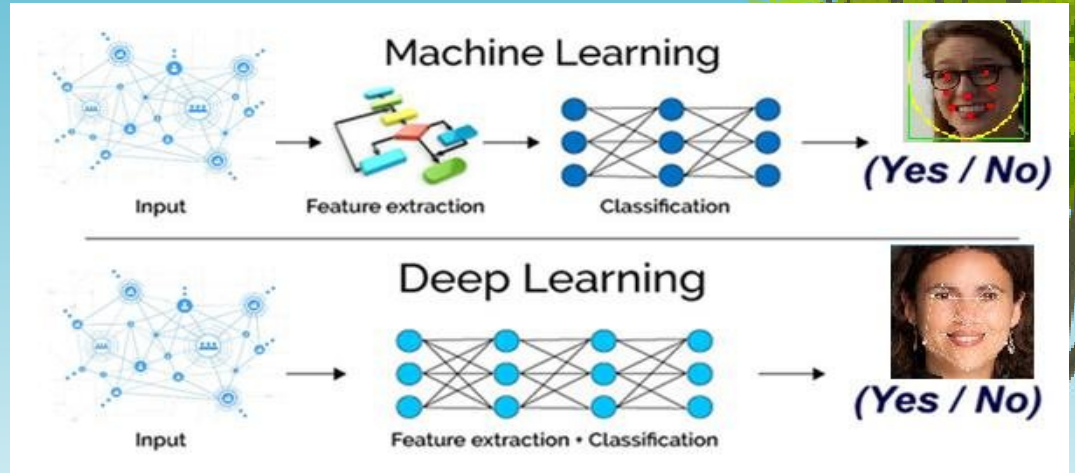
The Rise of Deep Learning

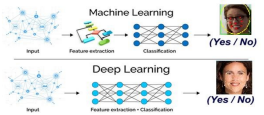




- GPUs have made it possible to expand accessibility to DL
- the CUDA toolkit from Nvidia has made things easier for researchers.

Quick Intro to a Giant Topic





- image source/credit
- ML feature extraction can be a huge undertaking, up to 80% of a project.
- DL attempts to automatically learn features that are most useful for a task from raw data.
- The nodes in a digraph are "neurons" or "units" in the DL/graph theory context.
- The neurons perform two steps. They calculate a "weighted sum" and pass the result through an "activation function" such as a rectifier activation function.
- These neurons or units that go through the rectifier function are called "RelUs" for short. Lot's of descriptive info in this one term!
- Depth of the GNN is measured by the number of connected layers.
- DL needs very large data sets for accurate feature determination. Data sets with lots of features are known as "high density".
- We humans interpret the features and output based on what we are trying to model.

Amazing Training and Tools Available

- There is a ton of information suddenly. Books, papers, code, etc.
- Folks are very helpful, positioning themselves as experts.
- super helpful videos like this one
- The Google Machine Learning Crash Course is free with tons of information.

- There is a ton of information suddenly. Books, papers, code, etc.
- Folks are very helpful, positioning themselves as experts.
- super helpful videos like this one
- The Google Machine Learning Crash Course is free with tons of information.

- Google Deep Learning Container Images
- Continuous Machine Learning (CML) Project
- Kaggle and shared Jupyter Notebooks

•

•

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

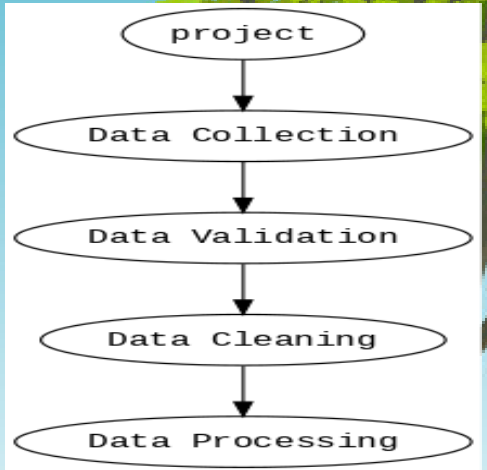
So Now What?

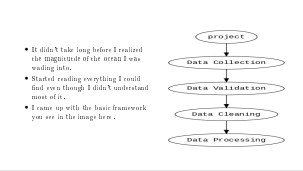
Sources and Citations

- Now I would like to talk a bit about the shape of the project.

What Have I Gotten Myself Into?

- It didn't take long before I realized the magnitutde of the ocean I was wading into.
- Started reading everything I could find even though I didn't understand most of it.
- I came up with the basic framework you see in the image here.





- Repetition can be a slow and painful way to learn.
- Wasn't even sure what questions to ask. Slow going at first.

Yak Shaving, Side Quests, Endless Rabbit Holes

- Makefiles and GNU Autotools
- NVIDA Jetson Nano as cluster nodes
- SLURM cluster scheduler
- OpenMPI for parallel builds
- Docker and Containers
- k8s and Rancher k3s
- Data Version Control dvc.org
- Storing/accessing data in GCP buckets
- Continuous Machine Learning cml.dev
- Internal Pypi and Debian/Raspbian mirror (used too much bandwidth on home connection)

- Makefiles and GNU Autotools
- NVIDA Jetson Nano as cluster nodes
- SLURM cluster scheduler
- OpenMPI for parallel builds
- Docker and Containers
- k8s and Rancher k3s
- Data Version Control dvc.org
- Storing/accessing data in GCP buckets
- Continuous Machine Learning cml.dev
- Internal Pypi and Debian/Raspbian mirror (used too much bandwidth on home connection)

- Wasn't sure exactly where to drop this slide in the order.
- Trying to show that there have been a TON of side quests.
- Some of these were useful, some led to spin off projects. A lot of this is bookmarked for later when I get some "spare time" haha.

Dot Data Collection

- A big barrier to entry was removed by the ability to output a Directed Graph from Terraform.
- Click for video

```
# Generate a PNG from Terraform
terraform graph | dot -Tpng > graph.png

# Generate vector graphic from Terraform
terraform graph | dot -Tsvg -o graph.svg
```

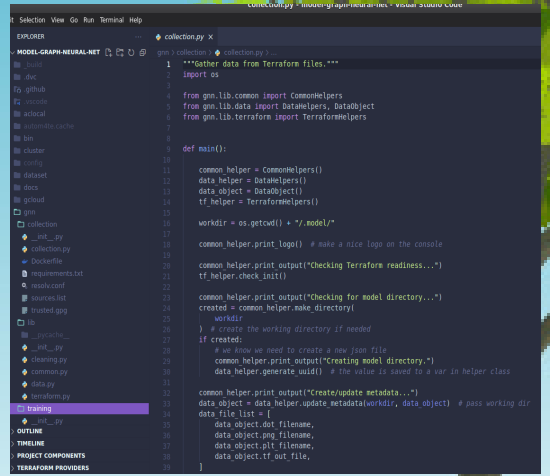

- A big barrier to entry was removed by the ability to output a Directed Graph from Terraform.
- Click for video

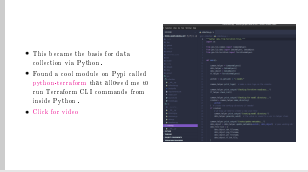
```
# Generate a PNG from Terraform
terraform graph | dot -Tpng > graph.png

# Generate vector graphic from Terraform
terraform graph | dot -Tsvg -o graph.svg
```

- Was pretty happy I could generate a PNG file. Super easy!
- Then I opened up the file and took a look at the nodes in the graph....

Python Data Collection

- This became the basis for data collection via Python.
- Found a cool module on Pypi called python-terraform that allowed me to run Terraform CLI commands from inside Python.
- Click for video

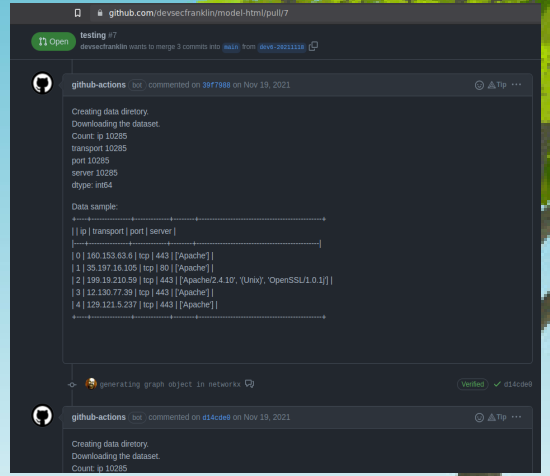


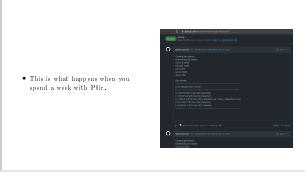


- Kind of a no brainer.
- The video is sped up 3x or so, but you can get the flavor of how the project looks from this.

Data Processing Side Quest

- This is what happens when you spend a week with P0lr.





- Spent a week with P0lr where we had a moderate case of machine learning fever.
- We had a direction but no destination.
- Watched Alpha Go movie, talked about a bunch of stuff, read some books and papers.
- wound up writing some code.
- there was a “HTML model” in there somewhere too

Data Storage - Kaggle

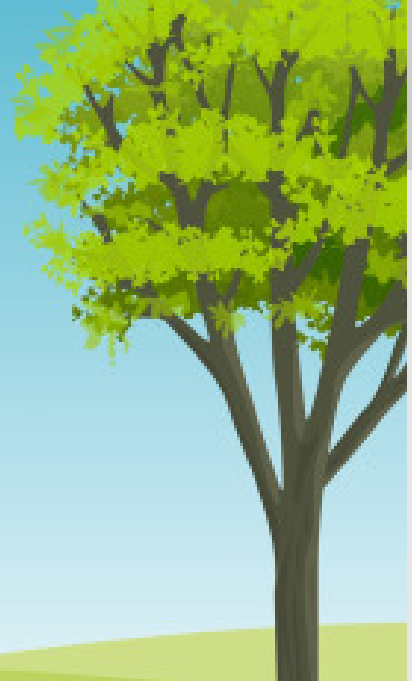
- What the heck is it?

- What the heck is it?

- What the heck is it?

Data Storage - Google Cloud

Data storage with GCP because it's (relatively) easy.



Data storage with GCP because it's (relatively) easy.

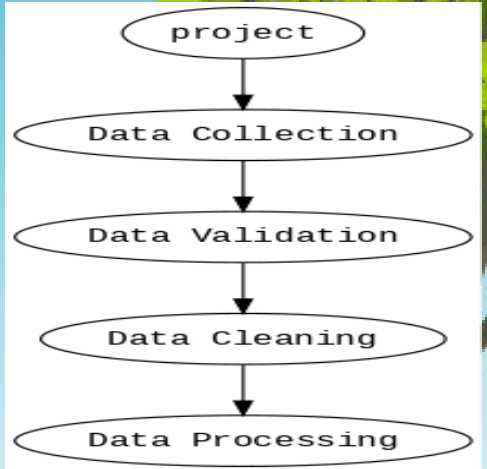
Data Storage - DVC

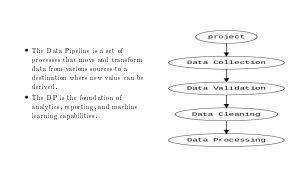
- Data storage and tagging using DVC
- there is a video on this page that explains

- Data storage and tagging using DVC
- there is a video on this page that explains

Data Pipeline

- The Data Pipeline is a set of processes that move and transform data from various sources to a destination where new value can be derived.
- The DP is the foundation of analytics, reporting, and machine learning capabilities.





- Source: Data Pipelines pocket reference p1-2

Introduction

What is Deep Learning, Exactly?

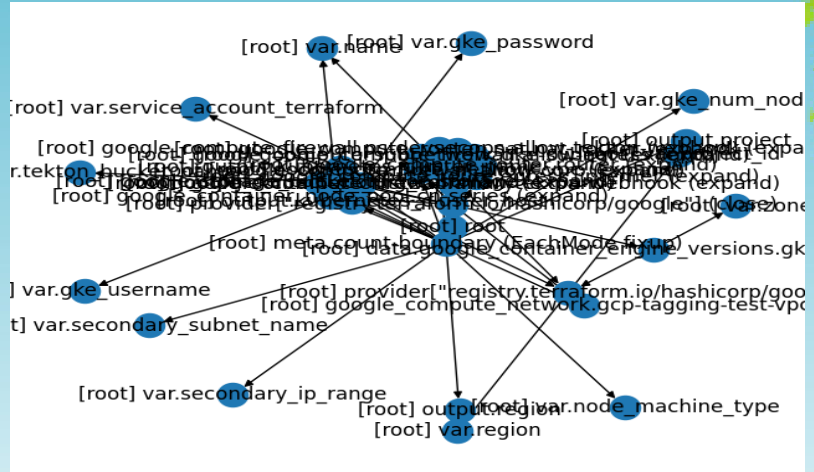
The Journey

Visualizations

So Now What?

Sources and Citations

Graphviz/Dot output



- This is the first thing I saw when I started converting the data.
- Was excited here since I was able to change the color of the nodes.
- Obviously this is not yet a usable result
- some video of data collection

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

Useful Intermediate Results

- Standardizing my data collection on JSON.
- Made some super cool functions for parsing nested JSON.
- Turned some of this time into money with cloud tools
- Importing JSON to Pandas dataframes.

So Now What?

- Standardizing my data collection on JSON.
- Made some super cool functions for parsing nested JSON.
- Turned some of this time into money with cloud tools
- Importing JSON to Pandas dataframes.

- Tabular data in Pandas can be output in all kinds of formats.
- Pandas data frames can be the input for other Machine Learning tools and frameworks.

Next Steps for this Project

- The data collection problem (Don't have access to enough data!)
- Maybe have a "collection container" with a python/Flask RESTful API for folks to push data to? Or even better, scrape GH for public repos with Terraform? (Lots of data, but not all security infra)
- Maybe back to Kaggle to find some big data to operate on?
- See if I can get the training to use my personal GPU/TPU.

So Now What?

- The data collection problem (Don't have access to enough data!)
- Maybe have a "collection container" with a python/Flask RESTful API for folks to push data to? Or even better, scrape GH for public repos with Terraform? (Lots of data, but not all security infra)
- Maybe back to Kaggle to find some big data to operate on?
- See if I can get the training to use my personal GPU/TPU.

- Most of this work is relegated to my "free" time.
- Have to spend my days helping people with the cloud.

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

Sources and Citations

- Generate a bibtex

• Generate a bibtex

- Need to add a reading list on here
- Make a list of all the books and papers