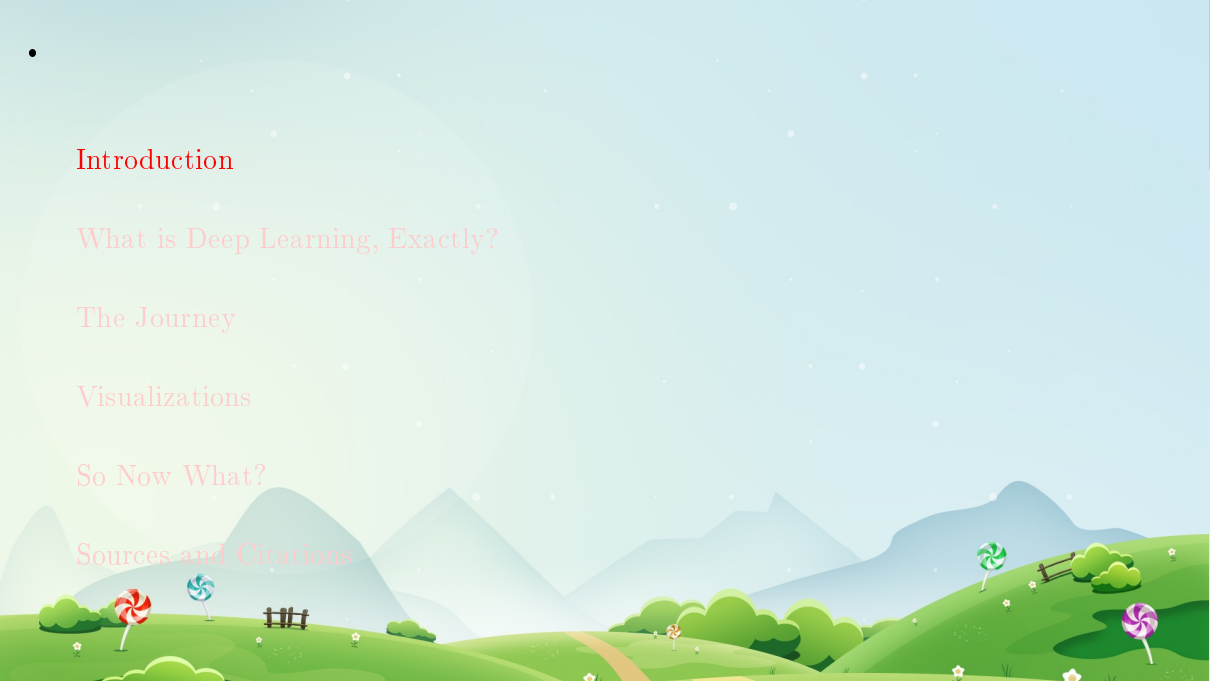


Learning About Deep Learning, and Maybe a Few Other Things

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

2022-05-01

└ Introduction

└ •

This first section is an Introduction to me and my project

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

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.



2022-05-01

Introduction

- 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.

- 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.



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](#)

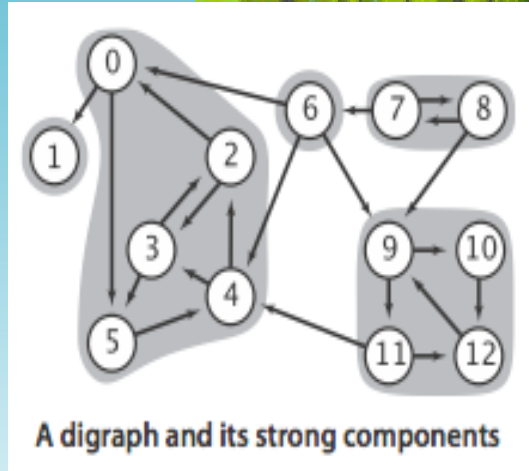
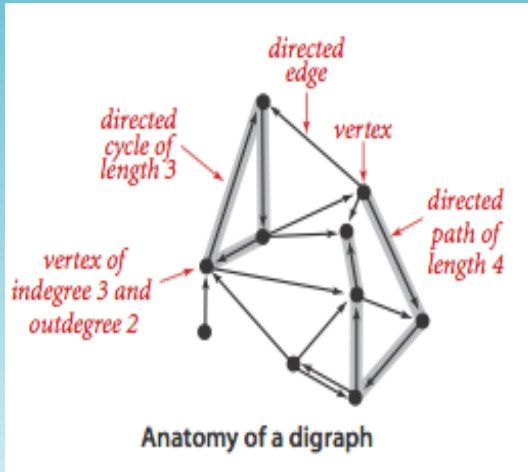
2022-05-01

└ Introduction

- 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?

- 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](#)

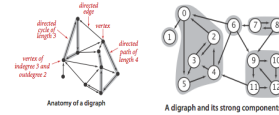
What's a DiGraph?



2022-05-01

Introduction

- 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

2022-05-01

└─ What is Deep Learning, Exactly?

└─ •

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

Introduction

What is Deep Learning, Exactly?

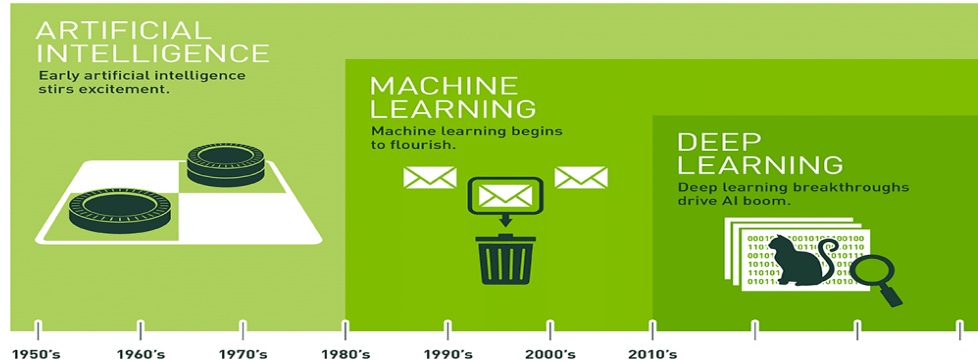
The Journey

Visualizations

So Now What?

Sources and Citations

The Rise of Deep Learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

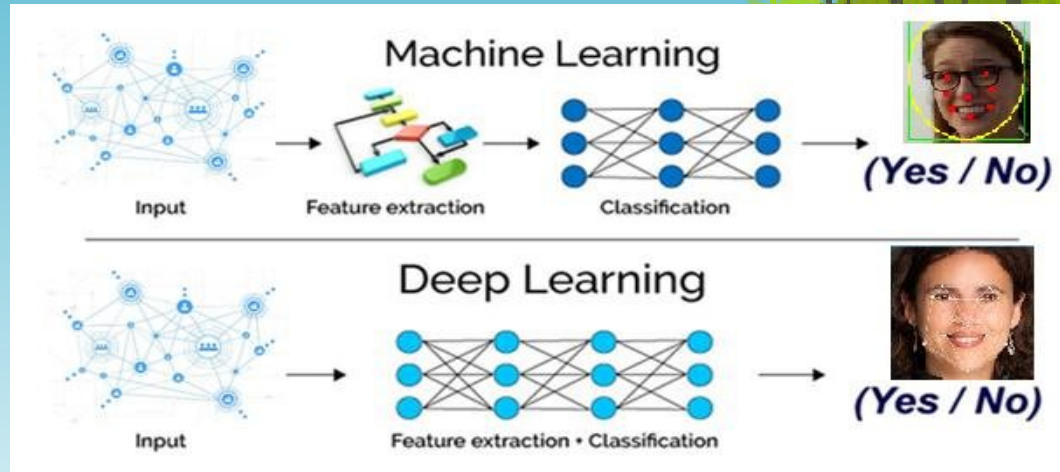
2022-05-01

What is Deep Learning, Exactly?

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

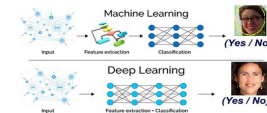


Quick Intro to a Giant Topic



2022-05-01

What is Deep Learning, Exactly?



- [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

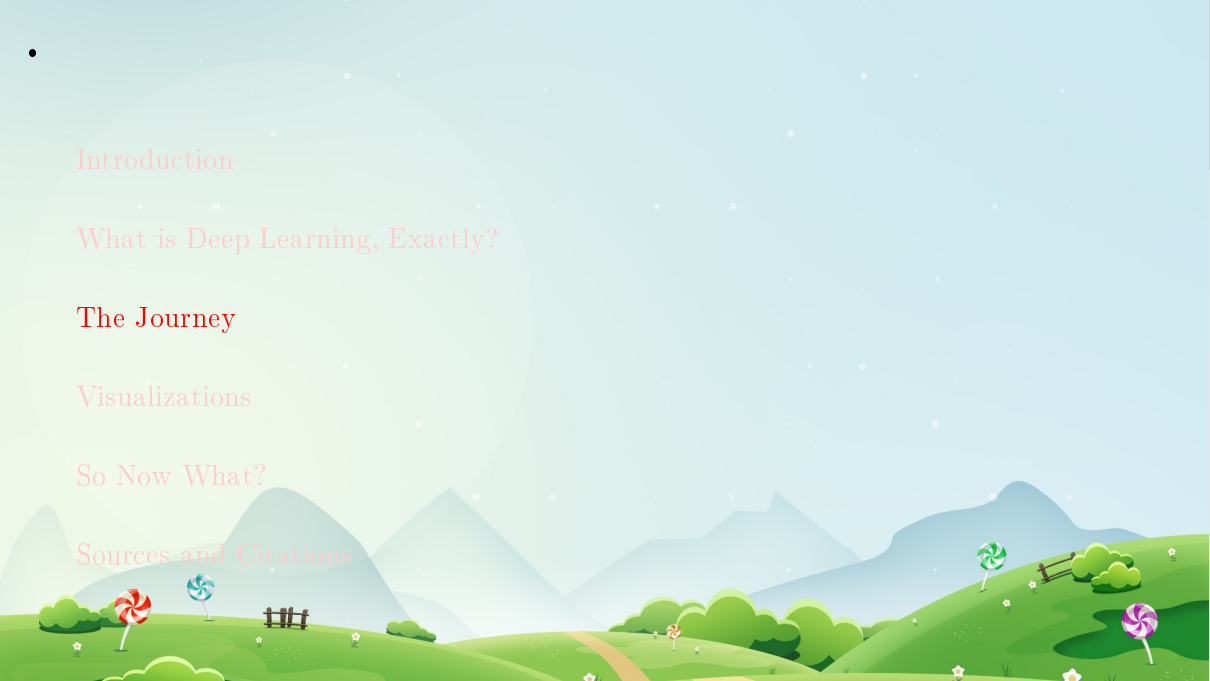
- There is a ton of information suddenly. Books, papers, code, etc.
- Folks are very helpful, positioning themselves as experts.
- [super helpful videos like this one](#)
- The [Google Machine Learning Crash Course](#) is free with tons of information.

2022-05-01

└─What is Deep Learning, Exactly?

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

- 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.



Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

2022-05-01

The Journey

•

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

Introduction

What is Deep Learning, Exactly?

The Journey

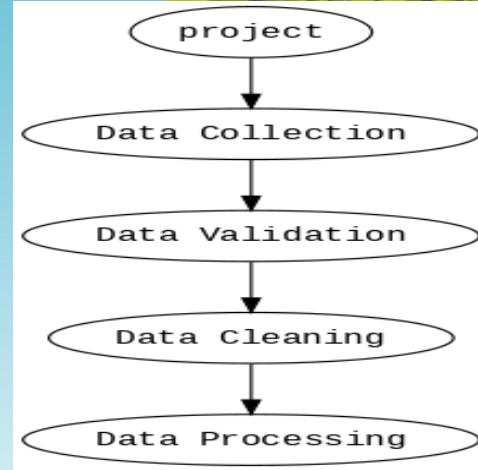
Visualizations

So Now What?

Sources and Citations

What Have I Gotten Myself Into?

- It didn't take long before I realized the magnitude 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.



2022-05-01

The Journey

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

- It didn't take long before I realized the magnitude 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.



Yak Shaving, Side Quests, Endless Rabbit Holes

- Makefiles and GNU Autotools
- NVIDIA 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)

2022-05-01

The Journey

- 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.

- Makefiles and GNU Autotools
- NVIDIA 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)

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
```

2022-05-01

The Journey

- 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....

• A big barrier to entry was removed by the ability to output a Directed Graph from Terraform.

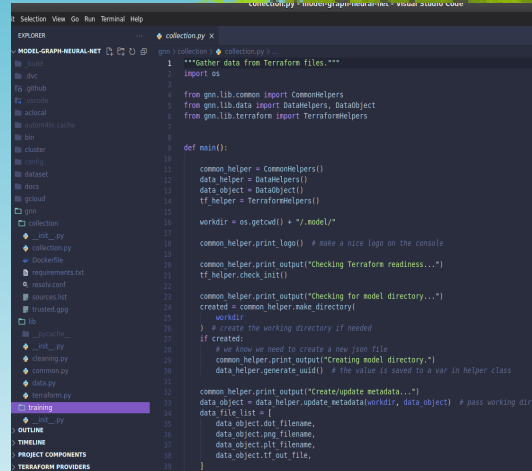
• [Click for video](#)

```
# generate a png from terraform
terraform graph | dot -Tpng > graph.png

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

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](#)



```
1 """Gather data from Terraform files."""
2 import os
3
4 from gnn.lib.common import CommonHelpers
5 from gnn.lib.data import DataHelpers, DataObject
6 from gnn.lib.terraform import TerraformHelpers
7
8
9 def main():
10
11     common_helper = CommonHelpers()
12     data_helper = DataHelpers()
13     data_object = DataObject()
14     tf_helper = TerraformHelpers()
15
16     workdir = os.getcwd() + "/.model/"
17
18     common_helper.print_logo() # make a nice logo on the console
19
20     common_helper.print_output("Checking Terraform readiness...")
21     tf_helper.check_init()
22
23     common_helper.print_output("Checking for model directory...")
24     created = common_helper.make_directory(
25         workdir
26     ) # create the working directory if needed
27     if created:
28         # we know we need to create a new json file
29         common_helper.print_output("Creating model directory.")
30         data_helper.generate_uuid() # the value is saved to a var in helper class
31
32     common_helper.print_output("Create/update metadata...")
33     data_object = data_helper.update_metadata(workdir, data_object) # pass working dir
34     data_file_list = [
35         data_object.dot_filename,
36         data_object.png_filename,
37         data_object.plt_filename,
38         data_object.tf_out_file,
39     ]
```

2022-05-01

The Journey

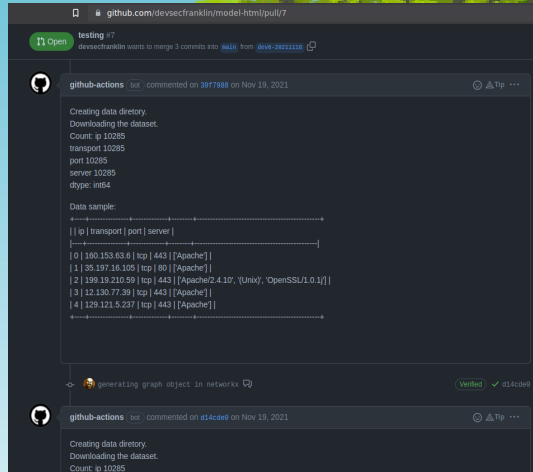
- Kind of a no brainer.
- The video is sped up 3x or so, but you can get the flavor of how the project looks from this.

- 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](#)



Data Processing Side Quest

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



The screenshot shows a GitHub pull request titled "testing #7" for the repository "devsecfranklin/model-html". The pull request is from branch "dev-2621118" to "main". The commit message is "devsecfranklin wants to merge 3 commits into main from dev-2621118". The pull request is open and has a green "Open" button. The pull request is for a file named "github-actions" and was commented on by "github-actions" on Nov 19, 2021. The code in the pull request shows a script for creating a data directory, downloading a dataset, and displaying a data sample. The data sample is a table with 5 columns: ip, transport, port, server, and dtype. The data is as follows:

	ip	transport	port	server	
[0]	160.153.63.6	tcp	443	[Apache]	
[1]	35.197.16.105	tcp	80	[Apache]	
[2]	199.19.210.59	tcp	443	[Apache/2.4.10', '(Unix)', 'OpenSSL/1.0.1j]]	
[3]	12.130.77.39	tcp	443	[Apache]	
[4]	129.121.5.237	tcp	443	[Apache]	

2022-05-01

The Journey

- 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

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



Data Storage - Kaggle

- What the heck is it?



2022-05-01

└ The Journey

- What the heck is it?

• What the heck is it?

Data Storage - Google Cloud

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



2022-05-01

└ The Journey

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

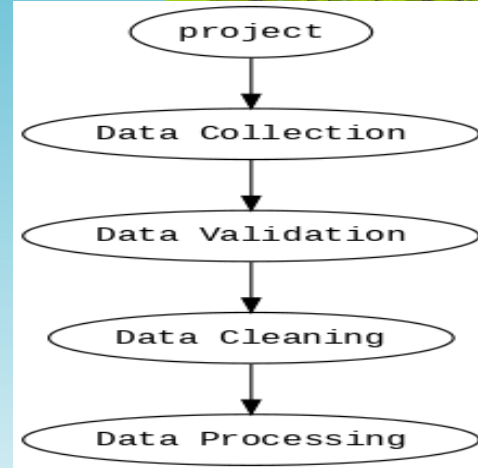
2022-05-01

└ The Journey

- 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.



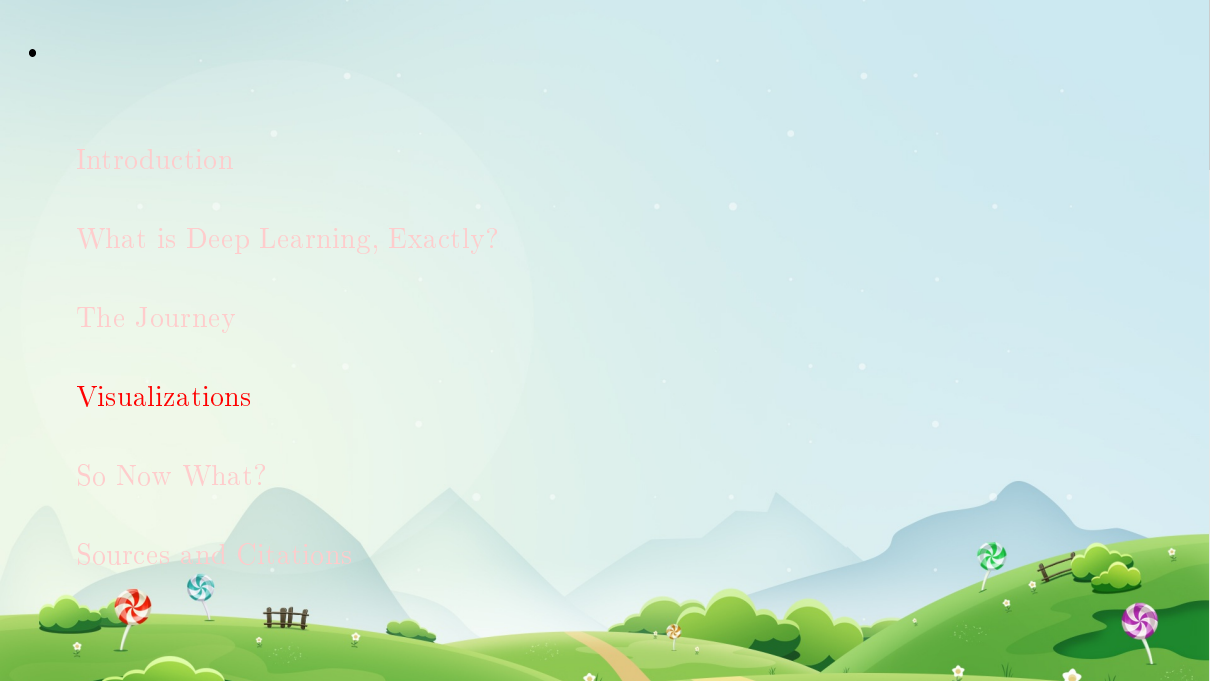
2022-05-01

The Journey

- Source: Data Pipelines pocket reference p1-2

- 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.





Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

2022-05-01

Visualizations

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

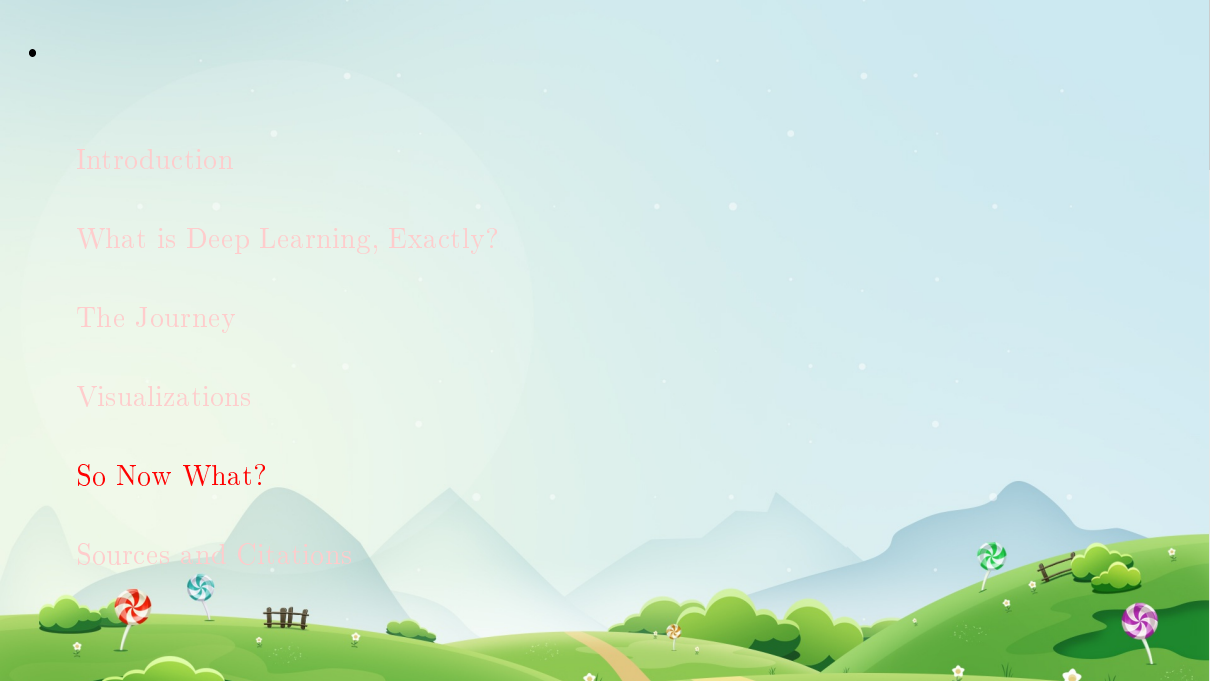
Sources and Citations

[illegible]

2022-05-01

└ Visualizations

- 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

2022-05-01

└─ So Now What?

└─ •

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.

2022-05-01

└ So Now What?

- 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.

- 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 .

Next Steps for this Project

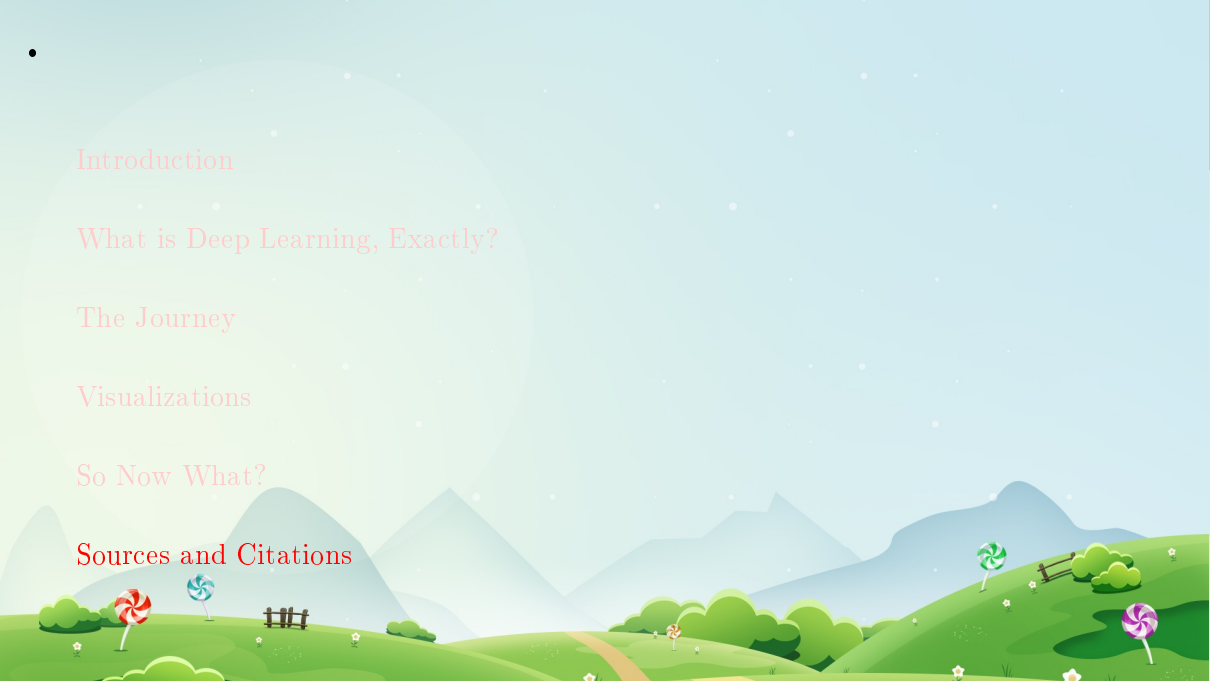
- The data collection problem (Don't have access to enough data!)
- Maybe have a “collection container” with a python/Flask RESTful API for folks to push data to? Or even better, scrape GH for public repos with Terraform? (Lots of data, but not all security infra)
- Maybe back to Kaggle to find some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

2022-05-01

So Now What?

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

- 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.



Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

2022-05-01

Sources and Citations

Introduction

What is Deep Learning, Exactly?

The Journey

Visualizations

So Now What?

Sources and Citations

Sources and Citations

- Generate a bibtex

2022-05-01

Sources and Citations

- Generate a bibtex

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