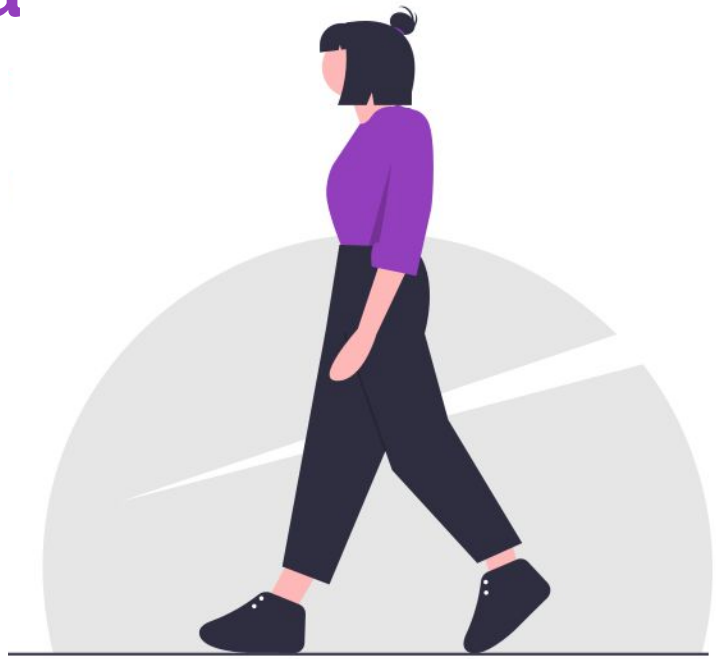
Introduction to Machine Learning and Web Scraping

Session 3, 19th Sep 2020

Agenda

1. Can we browse a web page programmatically?
2. What the heck is Machine Learning?
3. Can't we just write an algorithm? Why do we need data?
4. Is detecting Cat/Dog different from predicting new stock price tomorrow?



Let's Brush Python

Basics

Syntax

Comments

Variables

Data Types

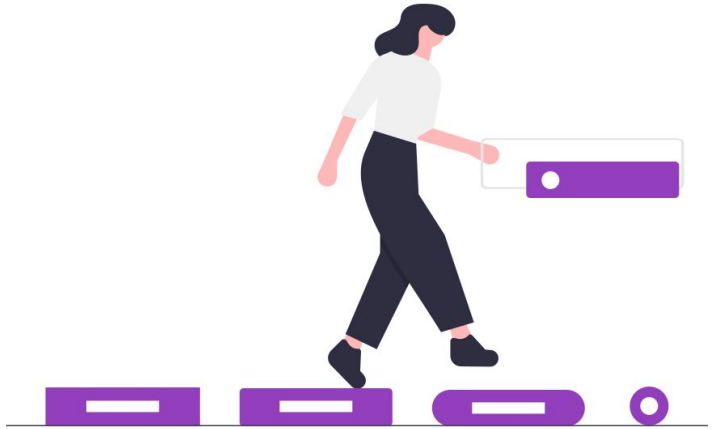
Operators

Loops

Conditions

Functions

Resource: **w3schools**



Numpy

Numpy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. A numpy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers. The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

The Python core library provided Lists. A list is the Python equivalent of an array, but is resizeable and can contain elements of different types.

Numpy

A common beginner question is what is the real difference here. The answer is performance. Numpy data structures perform better in:

- **Size** - Numpy data structures take up less space
- **Performance** - they have a need for speed and are faster than lists
- **Functionality** - SciPy and NumPy have optimized functions such as linear algebra operations built in.

Pandas

A dict is to a DataFrame as a bicycle is to a car. You can pedal 10 feet on a bicycle faster than you can start a car, get it in gear, etc, etc. But if you need to go a mile, the car wins.

For certain small, targeted purposes, a dict may be faster. And if that is all you need, then use a dict, for sure! But if you need/want the power and luxury of a DataFrame, then a dict is no substitute. It is meaningless to compare speed if the data structure does not first satisfy your needs.

Can we browse a web page programmatically?

Automating boring stuffs?

Web scraping is the process of extracting data on the web. With the right tools, anything that's visible to you can be extracted

Use Cases

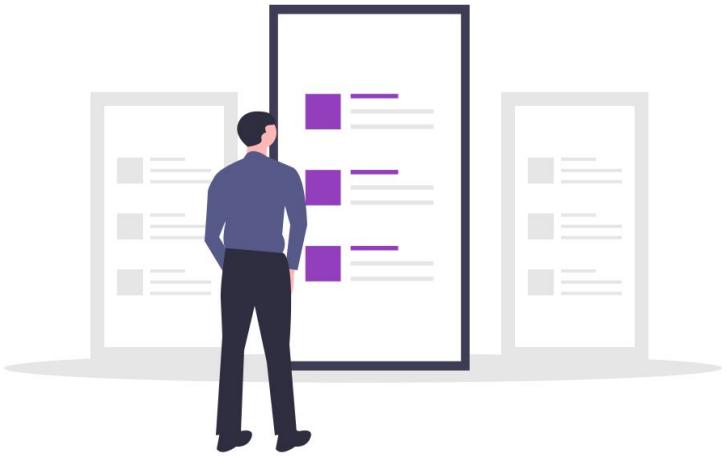
Price Comparison

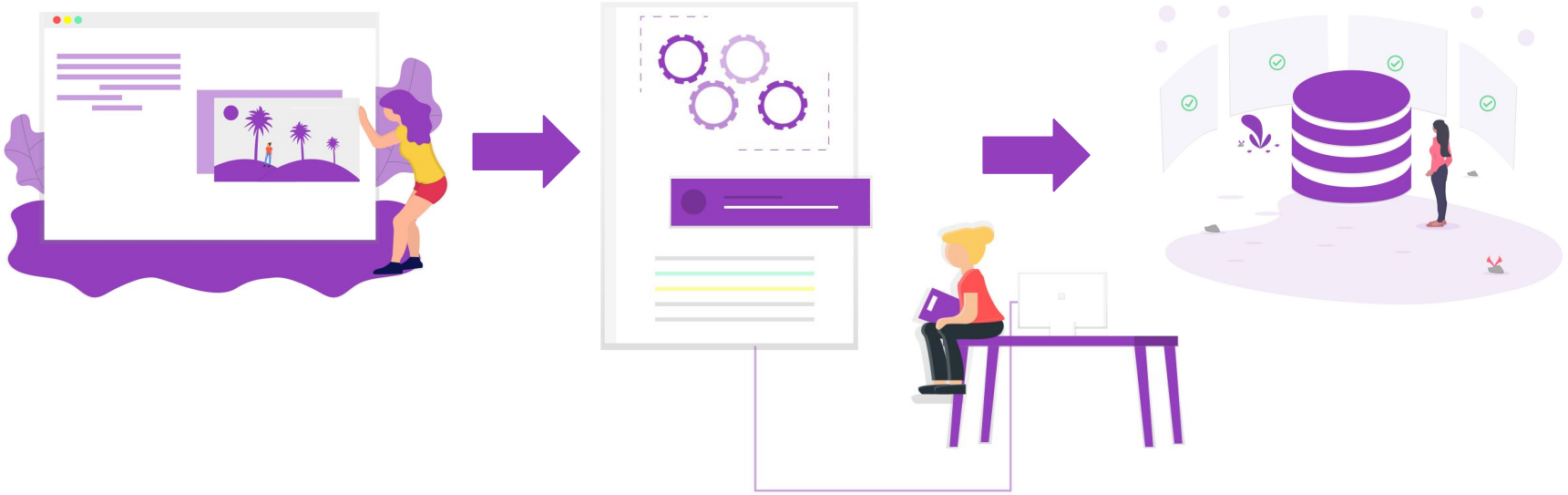
Email Address Gathering

Social Media Scraping

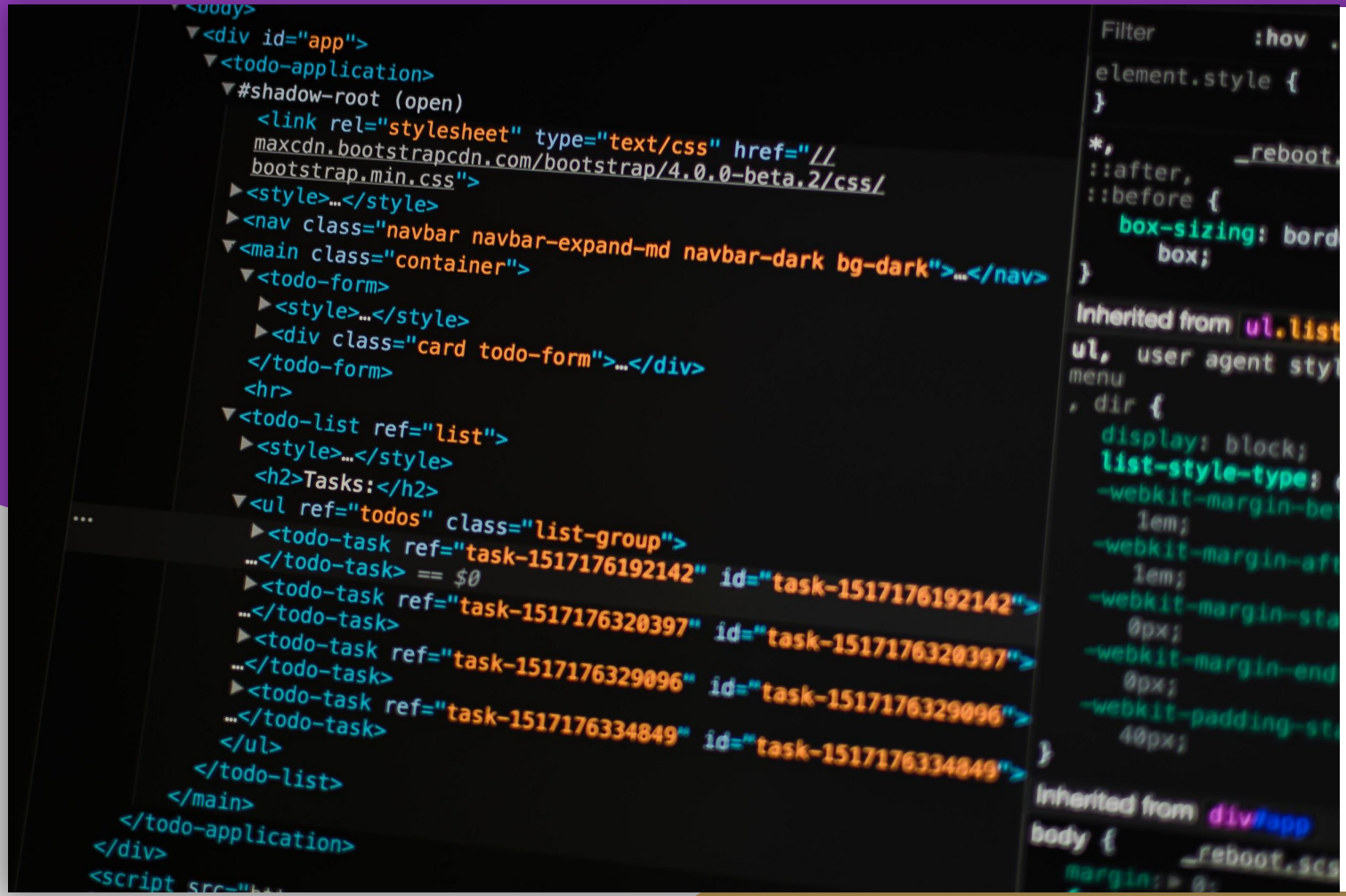
Research and Development

Job Listings





What is a Web Page?


▼<body>
▼
▼<todo-application>
▼#shadow-root (open)
<link rel="stylesheet" type="text/css" href="//
maxcdn.bootstrapcdn.com/bootstrap/4.0.0-beta.2/css/
bootstrap.min.css">
▶<style>…</style>
▶<nav class="navbar navbar-expand-md navbar-dark bg-dark">…</nav>
▼<main class="container">
▼<todo-form>
▶<style>…</style>
▶ …
</todo-form>
<hr>
▼<todo-list ref="list">
▶<style>…</style>
<h2>Tasks:</h2>
▼<ul ref="todos" class="list-group">
▶<todo-task ref="task-1517176192142" id="task-1517176192142">
…</todo-task> == $0
▶<todo-task ref="task-1517176320397" id="task-1517176320397">
…</todo-task>
▶<todo-task ref="task-1517176329096" id="task-1517176329096">
…</todo-task>
▶<todo-task ref="task-1517176334849" id="task-1517176334849">
…</todo-task>

</todo-list>
</main>
</todo-application>

<script src="
Filter
:hov
element.style {
}
*, _reboot.
::after,
::before {
box-sizing: bord
box;
}
Inherited from ul.list
ul, user agent styl
menu
, dir {
display: block;
list-style-type:
-webkit-margin-be
1em;
-webkit-margin-aft
1em;
-webkit-margin-sta
0px;
-webkit-margin-end
0px;
-webkit-padding-st
40px;
}
Inherited from div#app
body { _reboot.scs
margin: ▸ 0;

CV AI CHAMP

Demo

Let’s try to scrap some website

Machine Learning

1. What is **Machine Learning**?
2. What can we do with ML?
3. Types of ML
4. Common type of problem statements in ML
5. Popular Libraries

What is Machine Learning?

Machine Learning is Magic!

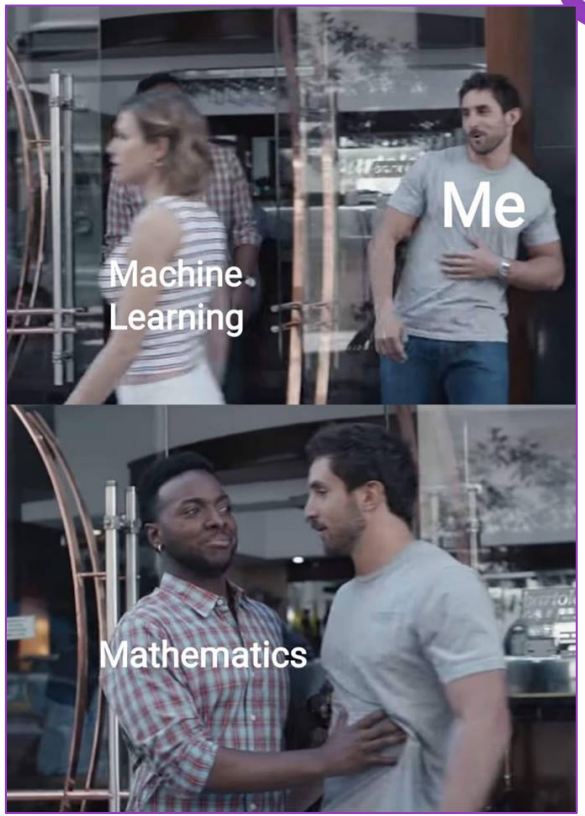
How many of you wanted to be magician?

Or at least do some magic?

Okay, I was kidding.

Machine Learning is Mathematics.

For me, Mathematics is magic.



How can computers learn from experience? Like we humans do?

→ We fail at many things. We say we've experienced something and try to not fail the next time.

→ Same goes with machines. And they have a lot of willpower, unlike me :D

→ They try for millions of times till they get things sorted out.

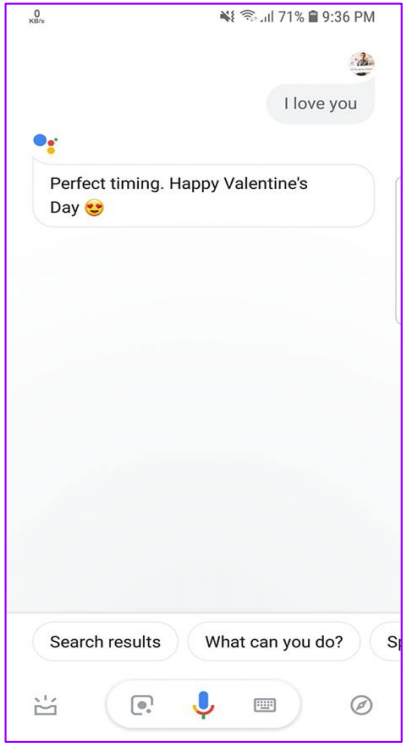
Where do machines get experience from?

One word.

DATA

What can we do with ML?

- Spam filtering
- Credit card fraud detection
- Digit recognition on checks, zip codes
- Detecting faces in images
- MRI image analysis
- Recommendation system
- Search engines
- Handwriting recognition
- Scene classification



AI, ML & DL

CV AI CHAMP

Artificial Intelligence

Any technique that enables computers to mimic human behavior

Machine Learning

Ability to learn without explicitly being programmed

Deep Learning

Extract patterns from data using neural networks

Most of the ML problems are one of the following

- → Regression
- → Classification
- → Clustering

Types of Machine Learning

→ Supervised

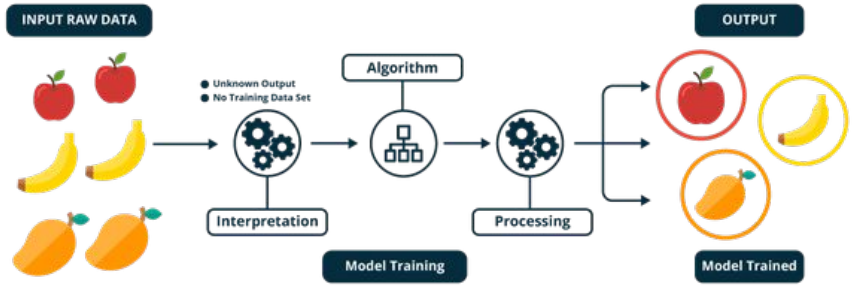
→ Unsupervised

→ Reinforcement Learning

→ Semi Supervised

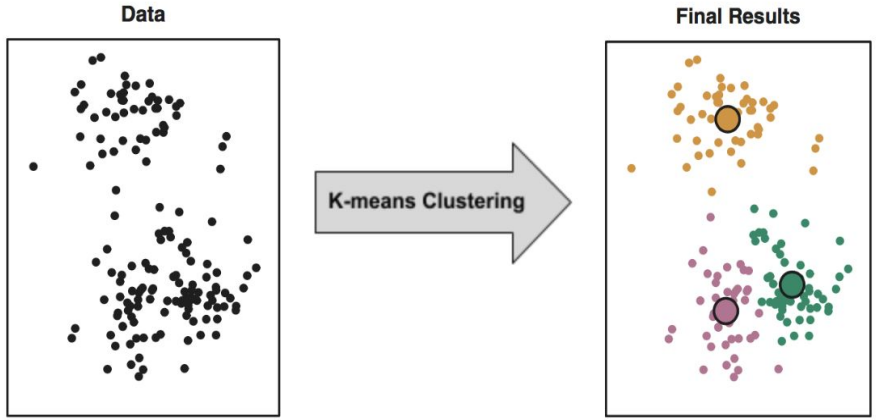
Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.



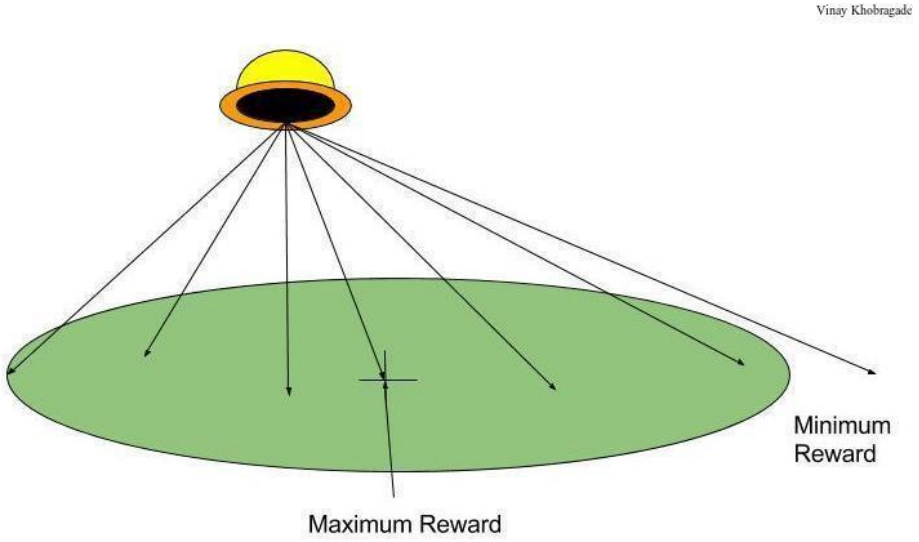
Unsupervised Learning

Unsupervised Learning is a class of Machine Learning techniques to find the patterns in data. The data given to unsupervised algorithm are not labelled, which means only the input variables(X) are given with no corresponding output variables.



Reinforcement Learning

Reinforcement learning is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.



Libraries

Libraries

numpy: Provides a fast numerical array structure and helper functions.

pandas: Provides a DataFrame structure to store data in memory and work with it easily and efficiently.

scikit-learn: The essential Machine Learning package in Python.

matplotlib: Basic plotting library in Python; most other Python plotting libraries are built on top of it.

seaborn: Advanced statistical plotting library.

Week Work

Week Work

Scrape **books.toscrape.com** and save the data to csv using pandas.

Note all the errors

Practice Numpy, Pandas

Read basic machine learning books, articles and create a glossary of terms

Participate in a kaggle competition, read available notebooks