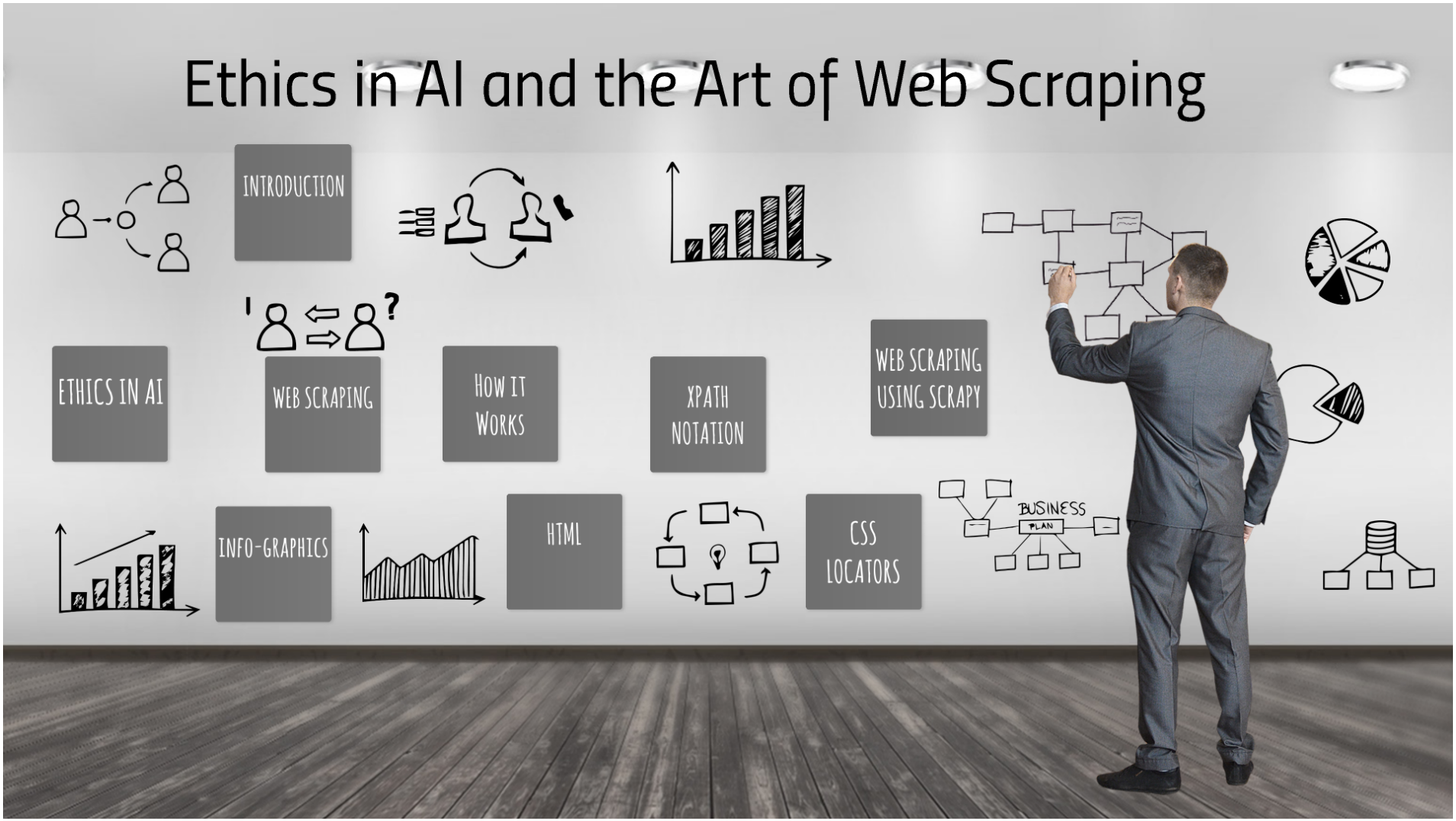
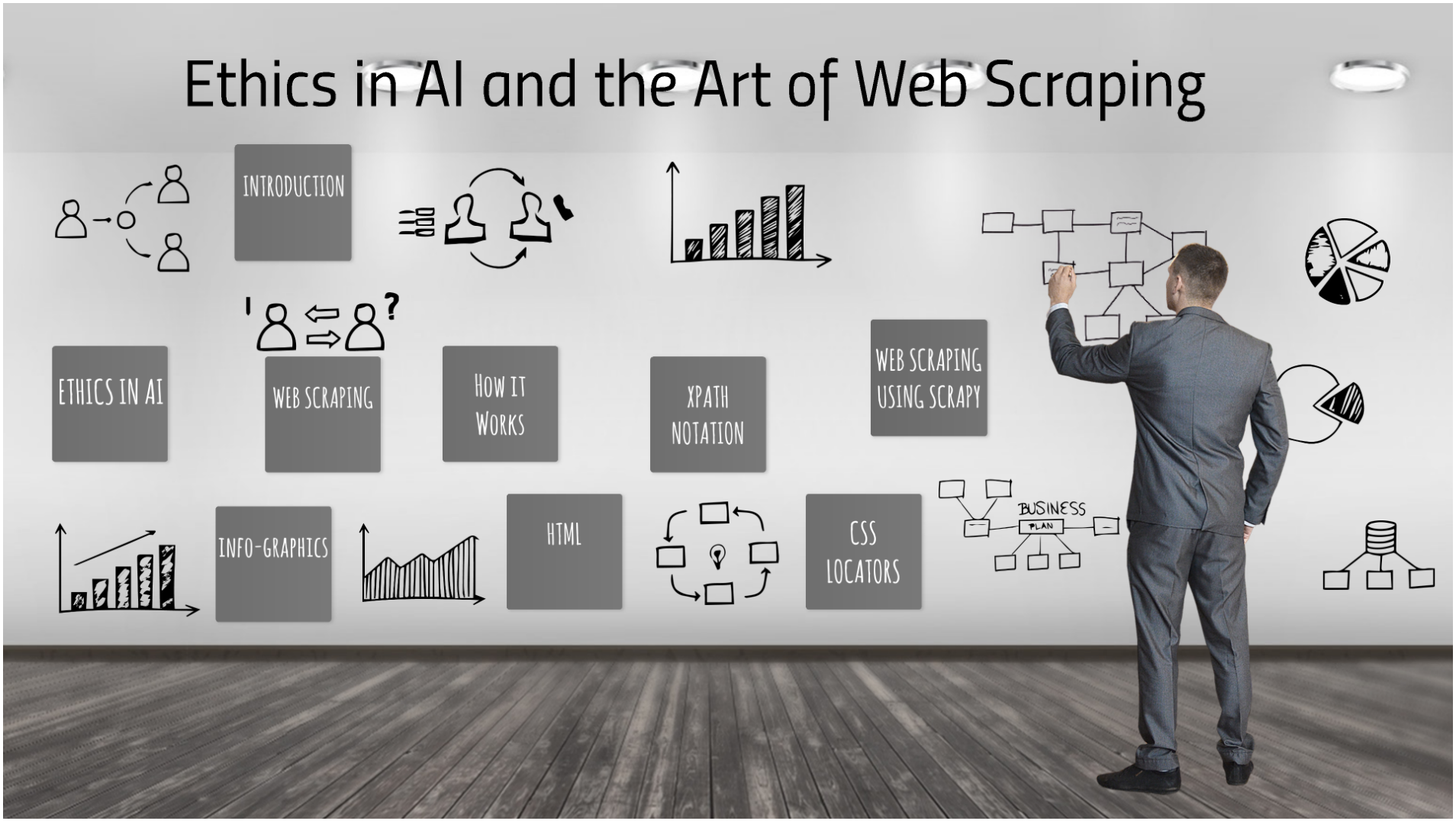

Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

INTRODUCTION

- Hi everyone this is Avirup
- I am a campus ambassador of bitgrit at SRM Institute of Science and Technology
- I am currently the co-founder and co-president of Data Science Community at SRM
- I am also working as the Associate Technical Lead at MSPC


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

What is Ethics

Ethics is defined as: The moral principles governing the behavior or actions of an individual or a group.

In other words, the “rules” or “decision paths” that help determine what is good or right.

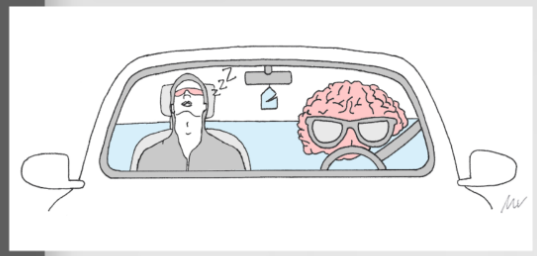
AI,ML etc are growing at a very rapid phase so its fundamental for a practitioner to understand where ethics comes in AI.

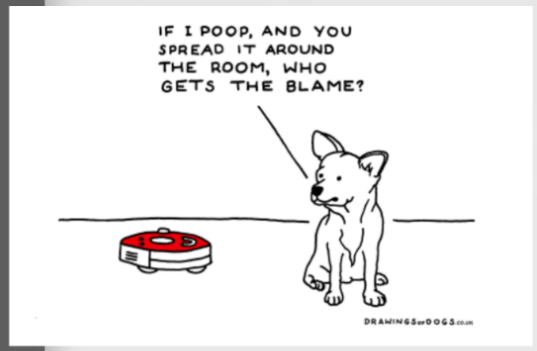
ETHICAL ISSUES IN AI

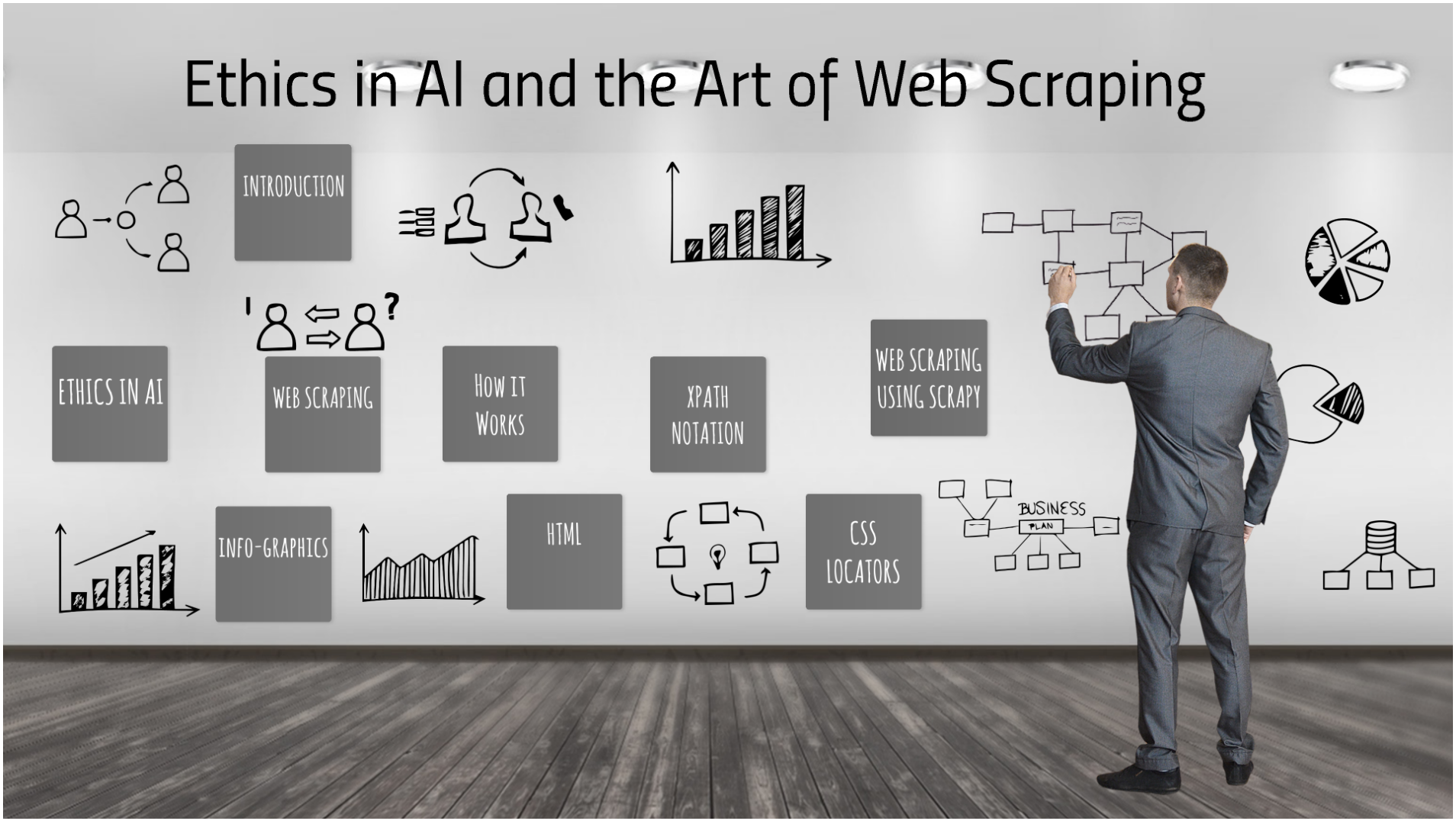
ETHICAL ISSUES PRESENT IN AI

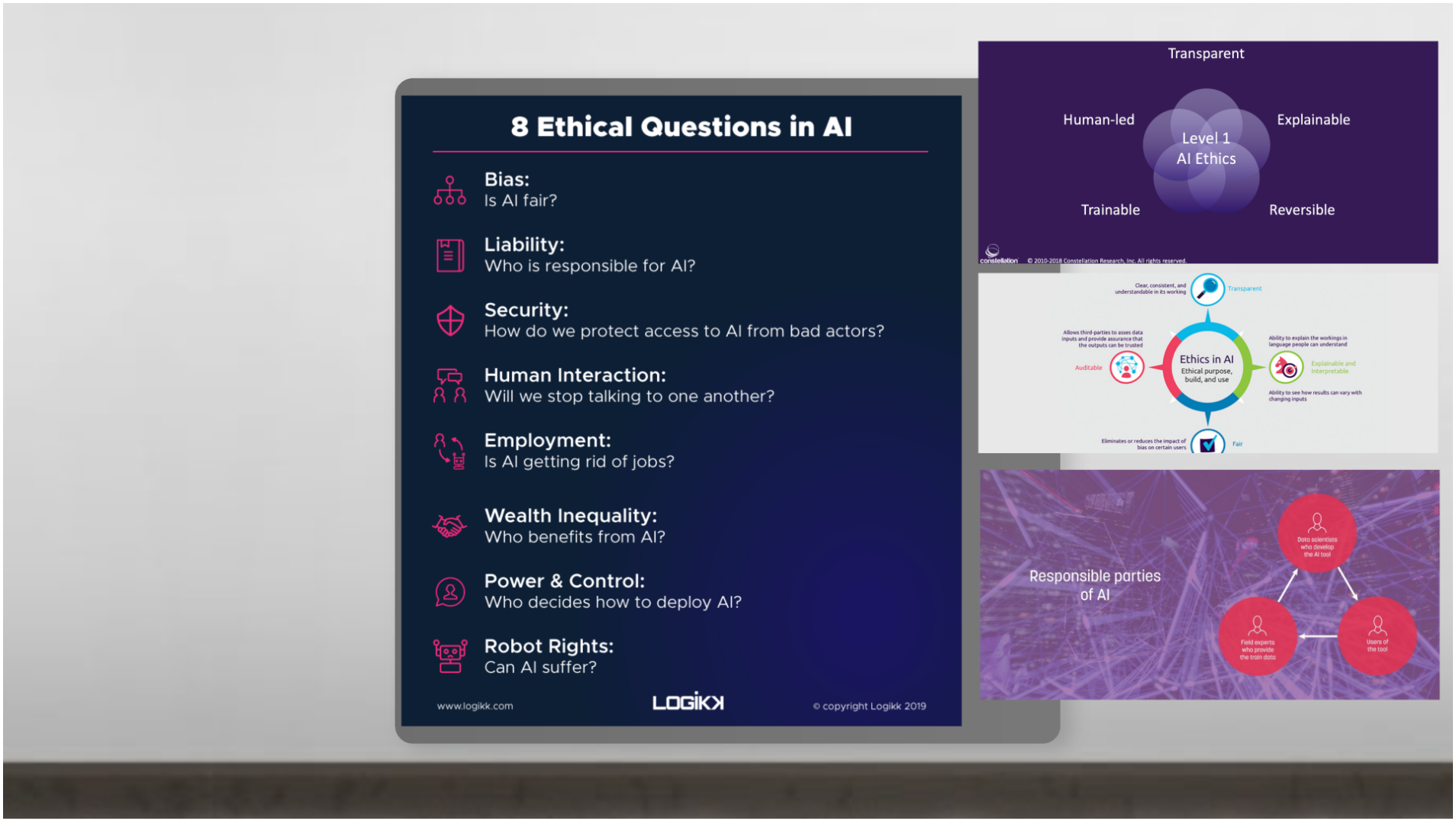
AI is growing at rapid phase and this has brought a number of ethical issues that we need to tackle with

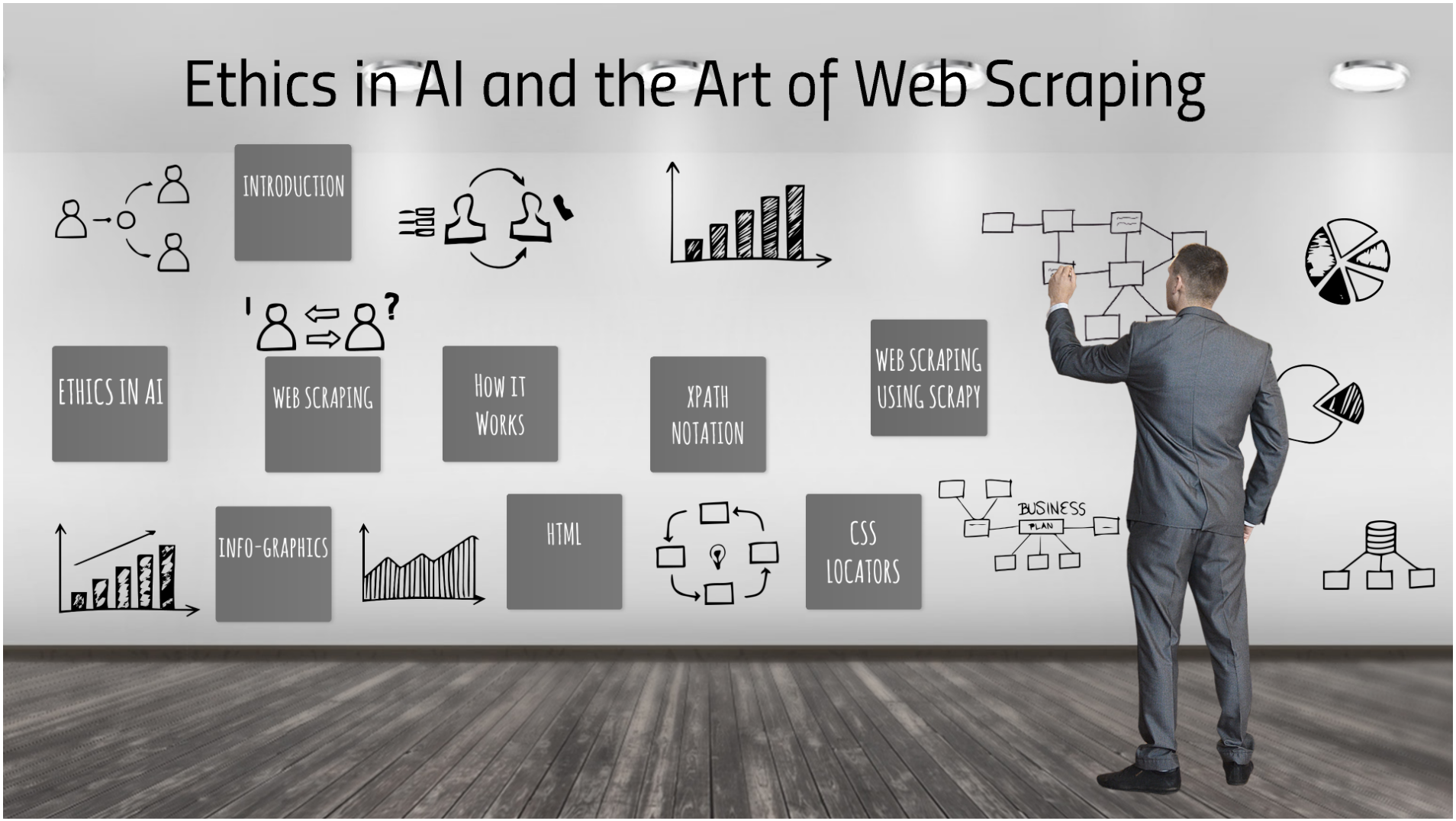
1.Unemployment
2.Inequality
3.Artificial Stupidity
4.Racist Bias
5.Evil Genies
6.Singularity
7.Robot Rights






Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN


8 Ethical Questions in AI
Bias:
Is AI fair?
Liability:
Who is responsible for AI?
Security:
How do we protect access to AI from bad actors?
Human Interaction:
Will we stop talking to one another?
Employment:
Is AI getting rid of jobs?
Wealth Inequality:
Who benefits from AI?
Power & Control:
Who decides how to deploy AI?
Robot Rights:
Can AI suffer?
www.logikk.com
LOGIKK
© copyright Logikk 2019
Transparent
Human-led
Explainable
Level 1
AI Ethics
Trainable
Reversible
constellation
© 2010-2018 Constellation Research, Inc. All rights reserved.
Clear, consistent, and understandable in its working
Transparent
Allows third-parties to asses data inputs and provide assurance that the outputs can be trusted
Auditable
Ethics in AI
Ethical purpose, build, and use
Ability to explain the workings in language people can understand
Explainable and Interpretable
Ability to see how results can vary with changing inputs
Eliminates or reduces the impact of bias on certain users
Fair
Responsible parties of AI
Data scientists who develop the AI tool
Field experts who provide the train data
Users of the tool


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

WEB SCRAPING

Web Scraping is a technique which is used to extract large amounts of data from a website and then store it in a local file

What are business looking for

WHAT A PERSON CAN GET FROM WEB SCRAPING

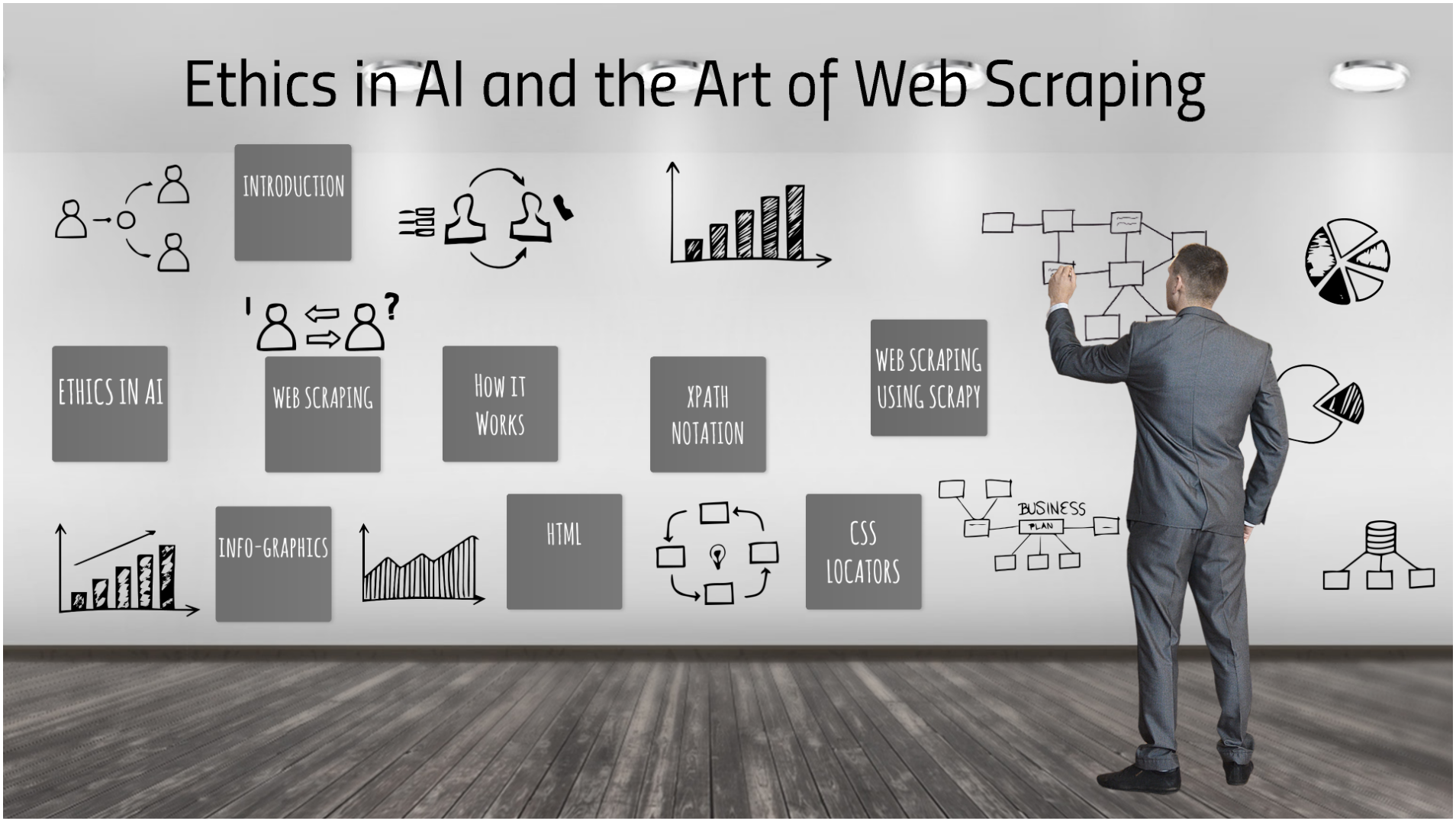
WHY DO BUSINESSES USE WEB SCRAPING

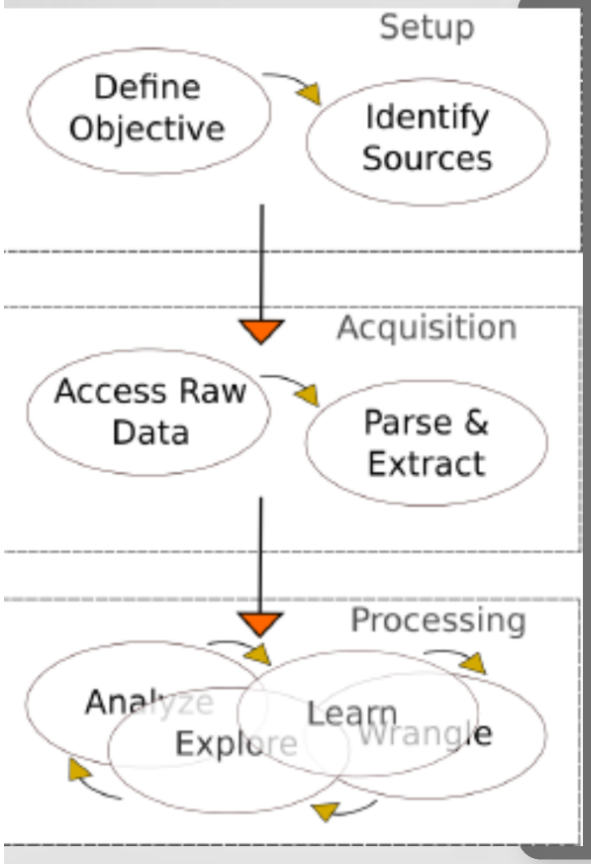
WEB SCRAPING ALLOWS BUSINESSES TO GET DATA REGARDING

- Comparing prices
- Satisfaction of customers
- Generating Potential Lead

WHAT A PERSON CAN GET FROM WEB SCRAPING

- Search for your favorite memes on your favorite sites.
- Scrape social site content looking for trendy topics
- Scrape cooking blogs looking for particular recipes, or recipe reviews.


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN



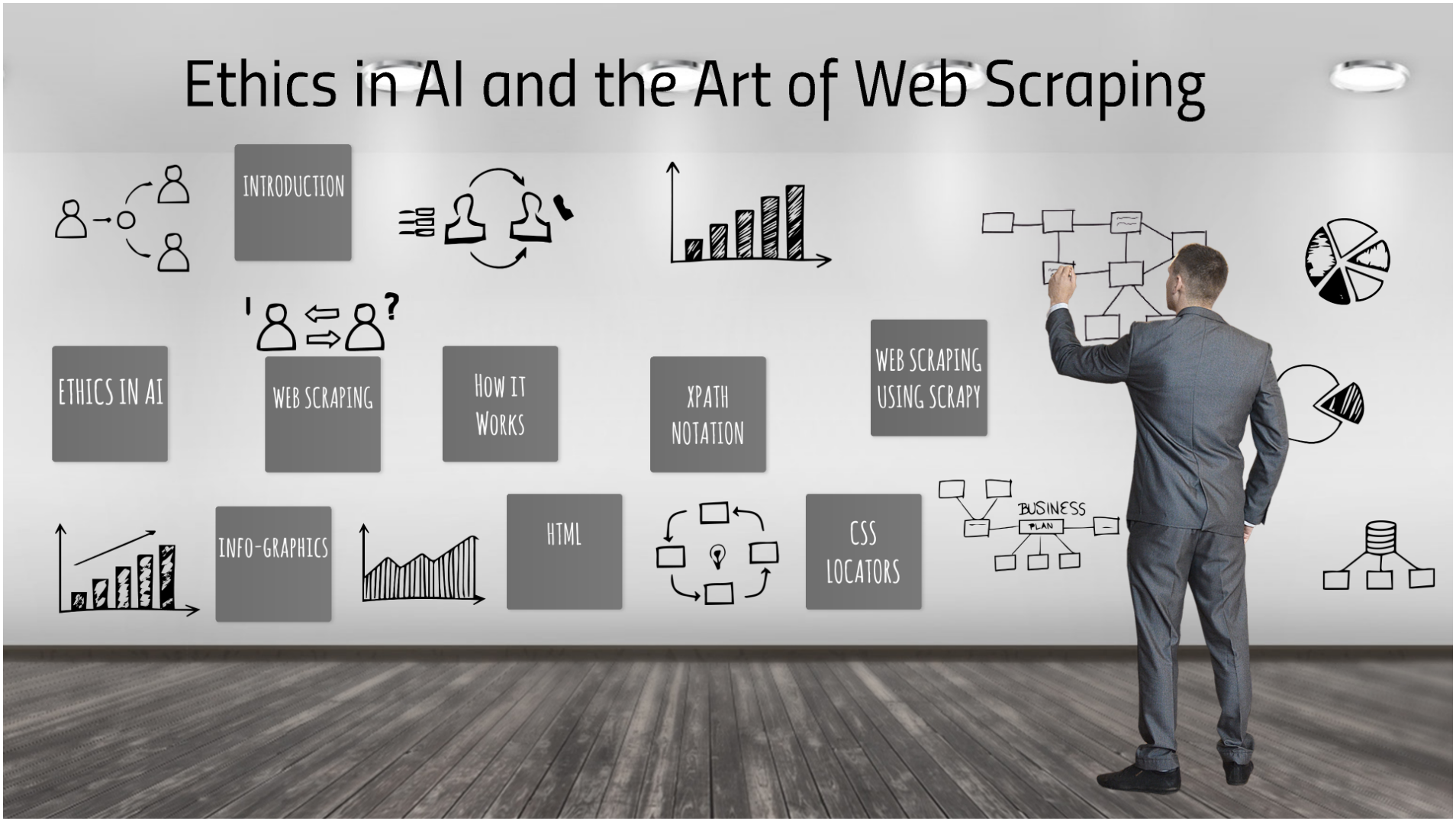
HOW DO WE DO WEB SCRAPING

WEB SCRAPING FOLLOWS THE THREE STEP PROCESS

- SETUP
- ACQUISITION
- PROCESSING

TOOLS WHICH ARE GOING TO BE USED FOR THIS SEMINAR ARE

- PYTHON
- SCRAPY
- REQUESTS


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

HTML

- HTML stands for Hyper Text Markup Language
- HTML is used to define the structure or the skeleton of the website
- Being a markup language it uses tags
- The current version of HTML is HTML5

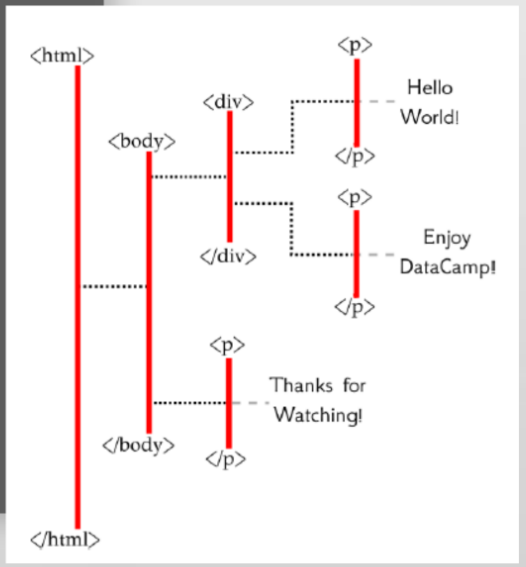
HTML TREE

HTML TAGS AND ATTRIBUTES

HTML TREE

- HTML can be interpreted as a tree like structure
- The root element is the main html tag
- So as we go down the tree we can go to the different elements present in the HTML document

```
<html>
    <body>
        <div>
            <p>Hello World!</p>
            <p>Enjoy DataCamp!</p>
        </div>
        <p>Thanks for Watching!</p>
    </body>
</html>
```

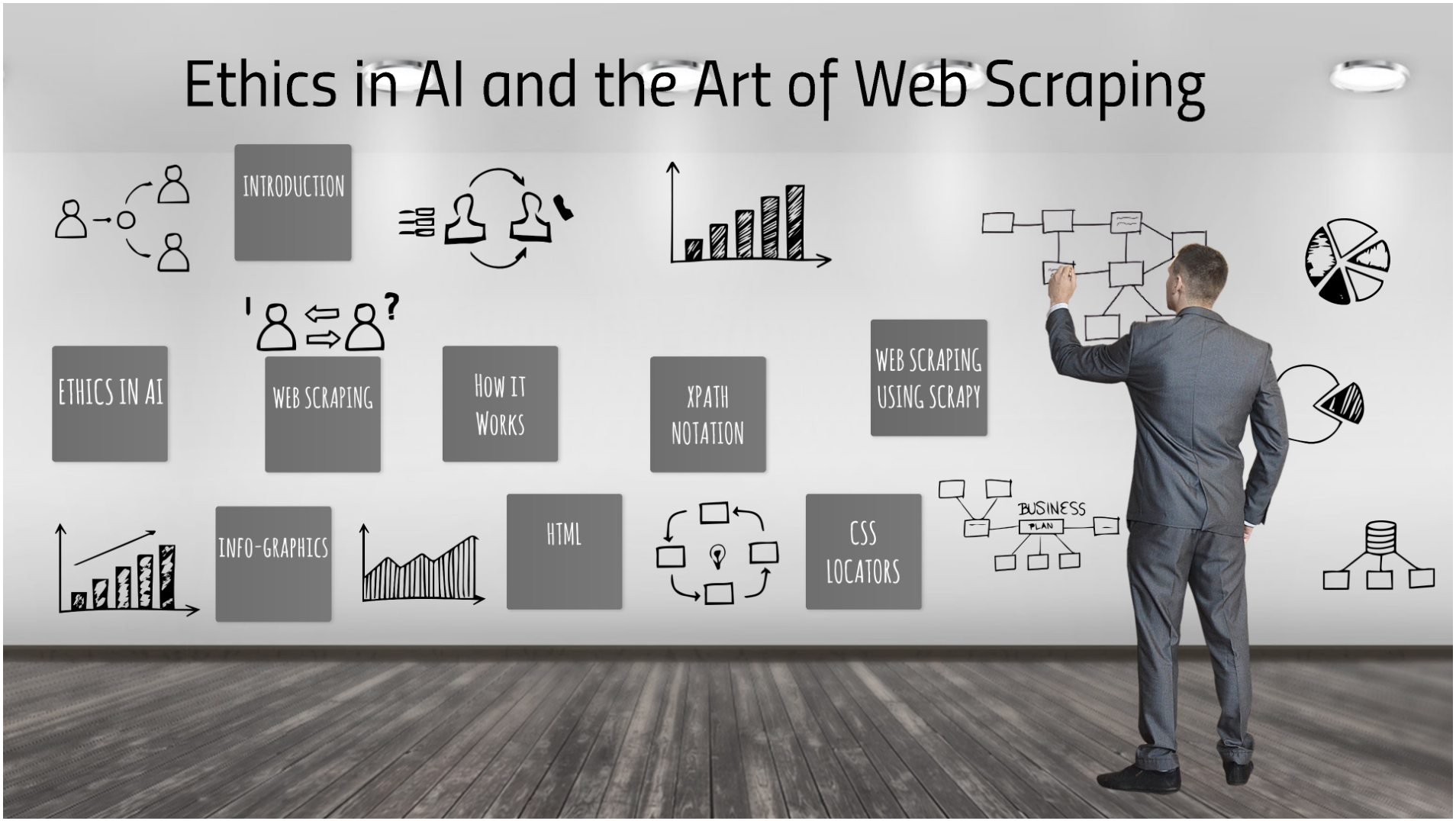


HTML TAGS AND ATTRIBUTES

```
<tag-name attrib-name="attrib info">
    ..element contents..
</tag-name>
```

- A html element is made up of the three parts which are a tag,attribute the data
- We have the start tag,end tag. We also have attributes which are like the special information for the element and the content or the data

```
<div id="unique-id" class="some class">
    ..div element contents..
</div>
```


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

XPATH

- XPATH stands for XML Path language
- XPATH is a query langugae which is used in order to extract elements from html and xml documents
- XPATH was developed by W3C and first appeared in 1998

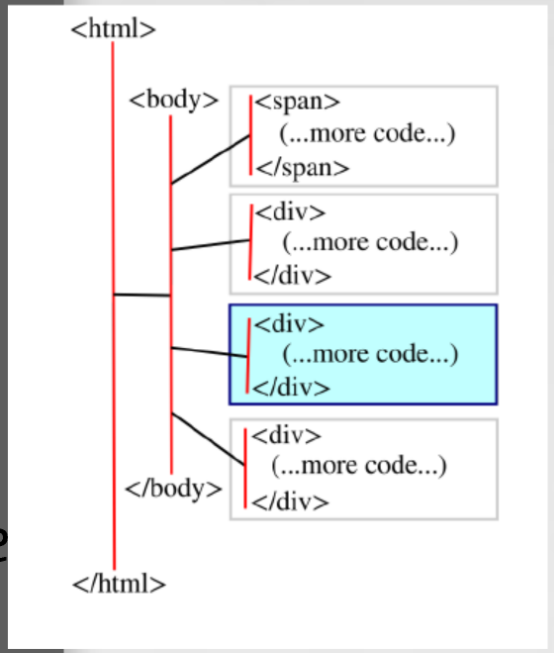
XPATH-1

XPATH-2

XPATH-3

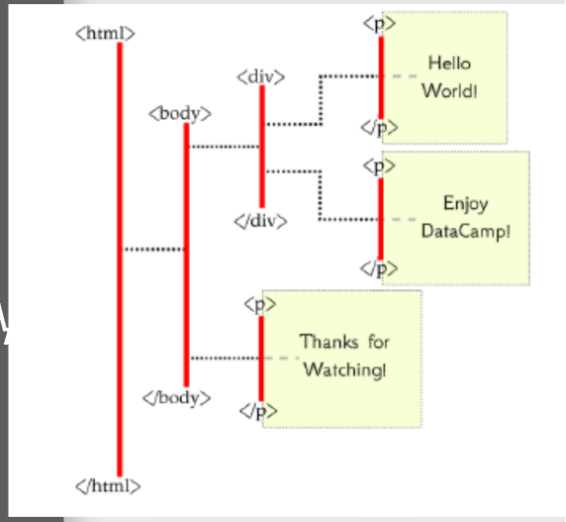
xpath = '/html/body/div[2]'

- Single forward slash is used to move from one generation to another
- tag-names between slashes give direction to which elements
- Brackets[] after a tag name tells us which sibling to choose

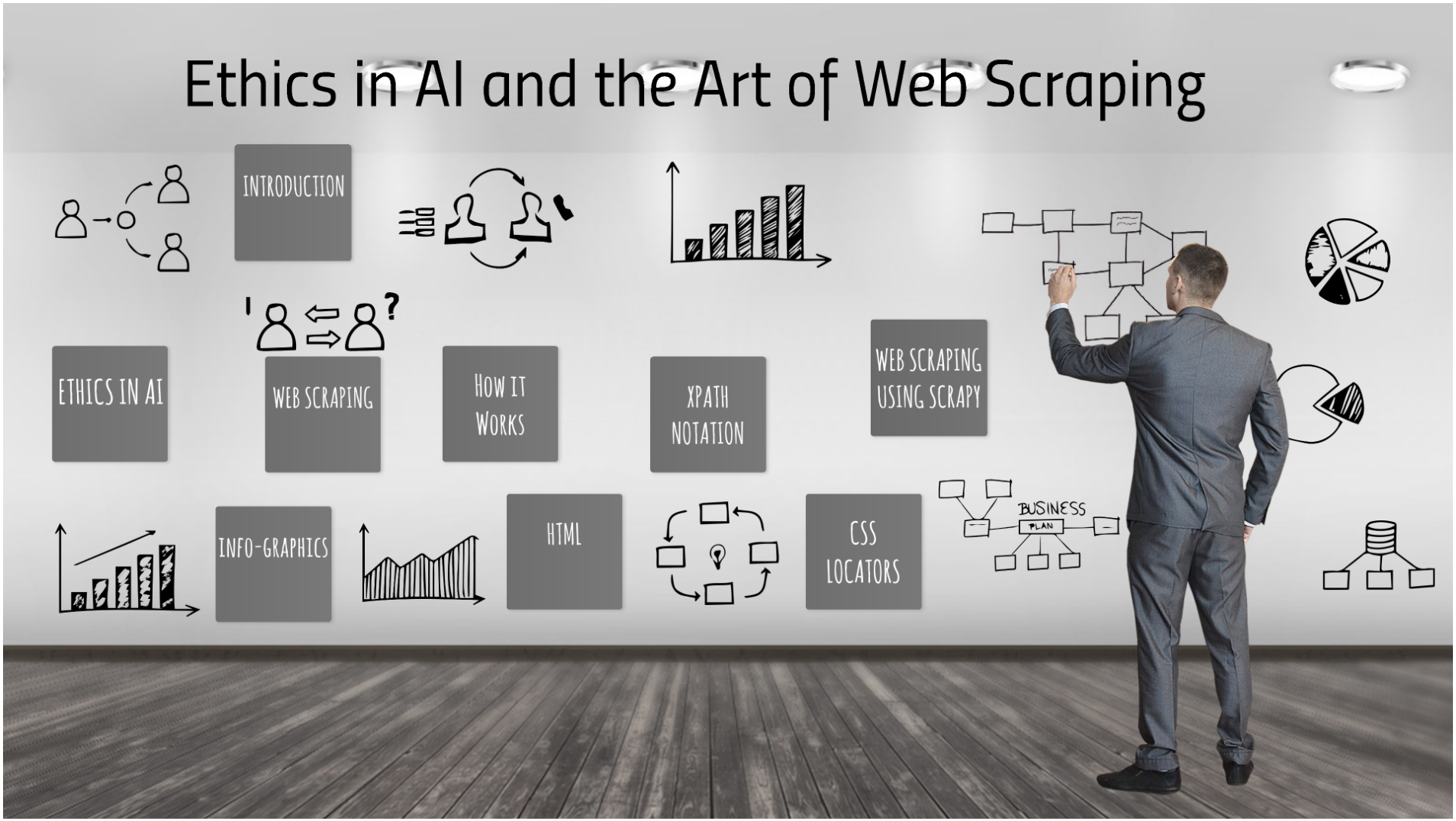


xpath = '//table'
Directs to all table elements within the entire HTML code

xpath = '//div[@class="body-class"]'
In order to extract html elements by their attributes we need to use the @. Similarly while accessing the html elements if we require attributes we can get using @



In order to get the data or the content out of the element we use the text() function.
In order to access any element we can even use * which is the wildcard character.


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

CSS LOCATORS/SELECTORS

- CSS locators/selectors are another method which are used to extract html elements out of an html document
- It follows the same procedure as xpath with syntactical differences and can be chained with xpath.

CSS-1

CSS-2

```
css = 'html > body div > p:nth-of-type(2)'
```

Some fundamental changes
/ gets replaced by >

// gets replaced by space

[N] gets replaced by :nth-of-type(N)

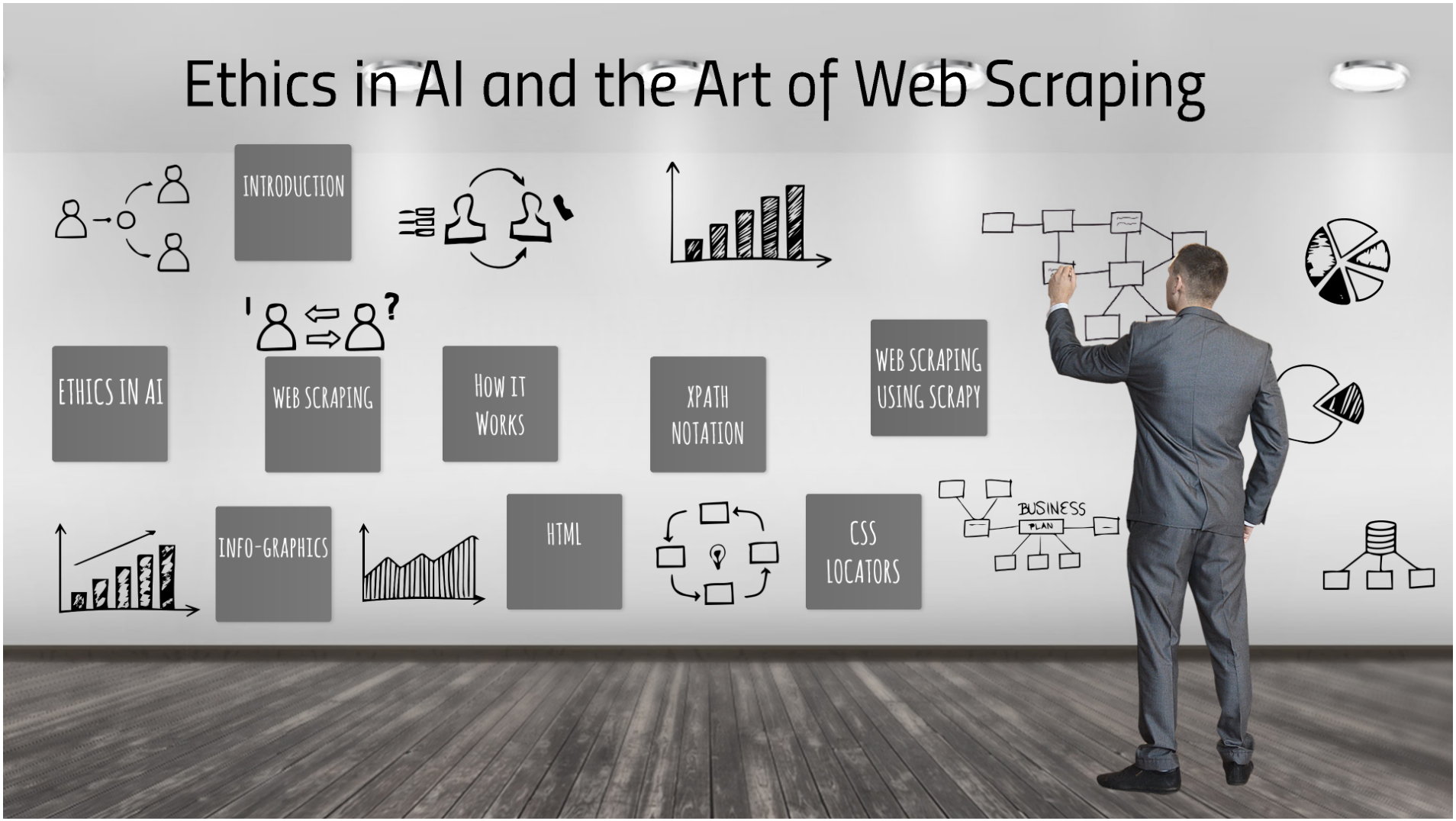
Attributes

In order to find an element by class we use the period(.) with the class value

In order to find an element by id we use the # with the id value

In order to get text from the html element we use ::text

In order to get the attribute we use ::attr(href)

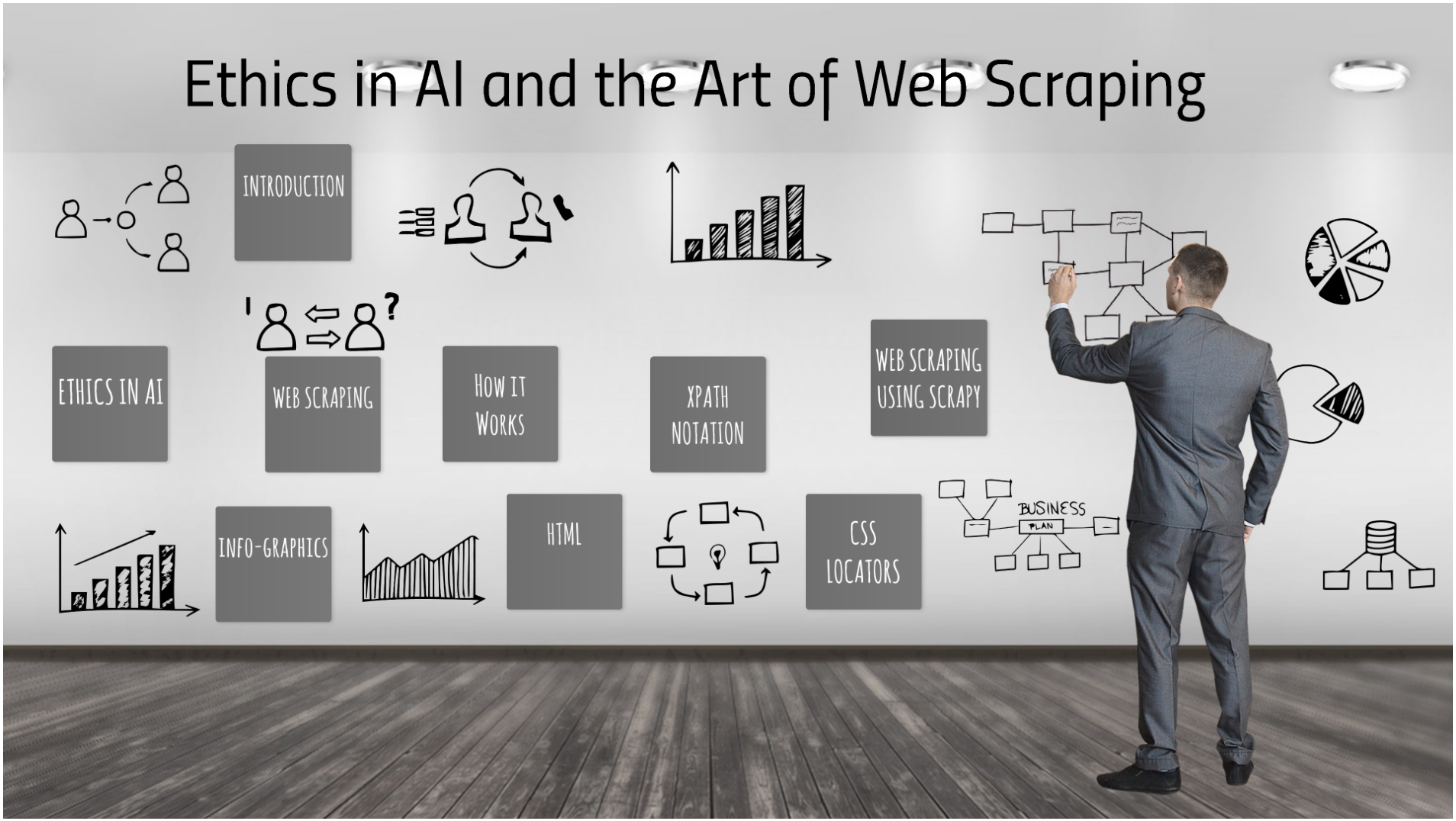

Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN

WEB SCRAPING USING SCRAPY

- Scrapy is an open source web crawling framework written in Python
- It's is currently maintained by ScrapingHub
- It's initial release happened in 2008 .

DATA EXTRACTION USING SELECTOR

```
import os
import requests
from scrapy import Selector
web_data=requests.get().conte
nt
sel=Selector(text=web_data)
data=sel.xpath('').extract()
for i in data:
    print(i)
```


Ethics in AI and the Art of Web Scraping
INTRODUCTION
ETHICS IN AI
WEB SCRAPING
HOW IT WORKS
XPATH NOTATION
WEB SCRAPING USING SCRAPY
INFO-GRAPHICS
HTML
CSS LOCATORS
BUSINESS
PLAN