In [30]:
```
import re
from pdfminer.high_level import extract_pages, extract_text
```

In [31]:
```
text = extract_text("/Users/elliewu/Downloads/Sample PDF File.pdf")
print(text)
```

```
Sample PDF File

This is an ordinary PDF File with some information.

Here are five names: Mike, Sara, Bob, John, Emma

Here are six numbers: 100, 200, 4310, 233, 544, 122

Here is a table full of data.

Name
Mike
Olivia
Bob
Sophia
Simon

Age
28
38
68
24
25

Job
Programmer
Accountant
Accountant
Lawyer
Programmer
```

In [39]:
```
#finding matches
pattern = re.compile(r"[a-zA-Z]+,{1}\s{1}")
matches = pattern.findall(text)
print(matches)
names = [n[:-2] for n in matches]
print(names)
```

```
['Mike, ', 'Sara, ', 'Bob, ', 'John, ']
['Mike', 'Sara', 'Bob', 'John']
```

In [40]:
```
#Extracting Images
import fitz #PyMuPDF
import PIL.Image #pillow
import io
```

In [48]:
```
pdf = fitz.open("/Users/elliewu/Downloads/Sample PDF File.pdf")
#1 image
counter = 1

for i in range(len(pdf)):
    page = pdf[i]
    images = page.get_images()
    for image in images:
        base_img = pdf.extract_image(image[0])
        image_data = base_img["image"]
        img = PIL.Image.open(io.BytesIO(image_data))
        extension = base_img["ext"]
```

```
        img.save(open(f"image{counter}.{extension}", "wb"))
        counter += 1
```

In [ ]: