1 Importing packages and the dataset

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib_venn as venn
from matplotlib_venn import venn2, venn2_circles, venn3, venn3_circles
import matplotlib.pyplot as plt
%matplotlib inline
```

executed in 675ms, finished 18:20:08 2021-06-07

In [2]:

```
df_venn_original=pd.read_csv('data_for_venn.csv')
df_venn_original.info()
```

executed in 79ms, finished 18:20:09 2021-06-07

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12379 entries, 0 to 12378
Data columns (total 36 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Unnamed: 0        12379 non-null  int64 
 1   COD_S11           12379 non-null  object
 2   GENDER            12379 non-null  object
 3   EDU_FATHER        12379 non-null  object
 4   EDU_MOTHER        12379 non-null  object
 5   OCC_FATHER        12379 non-null  object
 6   OCC_MOTHER        12379 non-null  object
 7   STRATUM           12379 non-null  object
 8   SISBEN            12379 non-null  object
 9   PEOPLE_HOUSE      12379 non-null  int64 
 10  INTERNET          12379 non-null  object
 11  TV                12379 non-null  object
 12  COMPUTER          12379 non-null  object
 13  WASHING_MCH       12379 non-null  object
 14  MIC_OVEN          12379 non-null  object
 15  CAR               12379 non-null  object
 16  DVD               12379 non-null  object
 17  FRESH             12379 non-null  object
 18  PHONE             12379 non-null  object
 19  MOBILE            12379 non-null  object
 20  REVENUE           12379 non-null  object
 21  JOB               12379 non-null  object
 22  SCHOOL_NAT        12379 non-null  object
 23  SCHOOL_TYPE       12374 non-null  object
 24  MAT_S11           12379 non-null  int64 
 25  CR_S11            12379 non-null  int64 
 26  CC_S11            12379 non-null  int64 
 27  BIO_S11           12379 non-null  int64 
 28  ENG_S11           12379 non-null  int64 
 29  Cod_SPro          12379 non-null  object
 30  UNIVERSITY        12379 non-null  object
 31  ACADEMIC_PROGRAM  12379 non-null  object
 32  G_SC              12379 non-null  int64 
 33  SEL               12379 non-null  int64 
 34  SEL_IHE           12379 non-null  int64 
 35  PROG_UNIV         12379 non-null  object
dtypes: int64(10), object(26)
memory usage: 3.4+ MB
```

2 Preprocessing the data and creating sets

In [3]:

```
df_venn_original=df_venn_original.drop(df_venn_original.columns[0], axis=1)
```

executed in 15ms, finished 18:20:09 2021-06-07

In [4]:

```
cols_to_drop=['PROG_UNIV','SCHOOL_TYPE', 'JOB','PEOPLE_HOUSE','EDU_FATHER','EDU_MOTHER','OCC_FATHER','OCC_MOTHER',
              'MAT_S11','CR_S11','CC_S11','BIO_S11','ENG_S11','Cod_SPro','UNIVERSITY','ACADEMIC_PROGRAM','G_SC','PROG_UNI
```

executed in 15ms, finished 18:20:09 2021-06-07

In [5]:

```
df_venn_original=df_venn_original.drop(cols_to_drop, axis=1)
```

executed in 15ms, finished 18:20:09 2021-06-07

In [6]:

```
df_venn_original=df_venn_original.astype({'SEL': 'object', 'SEL_IHE': 'object'})
```

executed in 14ms, finished 18:20:09 2021-06-07

In [7]:

```
df_venn_original.info()
```

executed in 15ms, finished 18:20:09 2021-06-07

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12379 entries, 0 to 12378
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   COD_S11      12379 non-null  object
 1   GENDER       12379 non-null  object
 2   STRATUM      12379 non-null  object
 3   SISBEN       12379 non-null  object
 4   INTERNET     12379 non-null  object
 5   TV           12379 non-null  object
 6   COMPUTER     12379 non-null  object
 7   WASHING_MCH  12379 non-null  object
 8   MIC_OVEN     12379 non-null  object
 9   CAR          12379 non-null  object
 10  DVD          12379 non-null  object
 11  FRESH        12379 non-null  object
 12  PHONE        12379 non-null  object
 13  MOBILE       12379 non-null  object
 14  REVENUE      12379 non-null  object
 15  SCHOOL_NAT   12379 non-null  object
 16  SEL          12379 non-null  object
 17  SEL_IHE      12379 non-null  object
dtypes: object(18)
memory usage: 1.7+ MB
```

In [8]:

```
df_venn_original.to_csv('data_for_venn_clean.csv', index=False)
```

executed in 62ms, finished 18:20:09 2021-06-07

In [9]:

```
df_venn=df_venn_original.copy()
```

executed in 15ms, finished 18:20:09 2021-06-07

In [10]:

```
ALL_set=set(df_venn['COD_S11'])
```

executed in 15ms, finished 18:20:09 2021-06-07

In [11]:

```
#print(len(list(df_venn.loc[df_venn.GENDER=='M']['COD_S11'])))
TV_set=set(df_venn.loc[df_venn.TV=='Yes']['COD_S11'])
Internet_set=set(df_venn.loc[df_venn.INTERNET=='Yes']['COD_S11'])
Computer_set=set(df_venn.loc[df_venn.COMPUTER=='Yes']['COD_S11'])
Washing_machine_set=set(df_venn.loc[df_venn.WASHING_MCH=='Yes']['COD_S11'])
Microwave_set=set(df_venn.loc[df_venn.WASHING_MCH=='Yes']['COD_S11'])
Car_set=set(df_venn.loc[df_venn.CAR=='Yes']['COD_S11'])
DVD_set=set(df_venn.loc[df_venn.DVD=='Yes']['COD_S11'])
Fresh_set=set(df_venn.loc[df_venn.FRESH=='Yes']['COD_S11'])
Phone_set=set(df_venn.loc[df_venn.PHONE=='Yes']['COD_S11'])
Mobile_set=set(df_venn.loc[df_venn.MOBILE=='Yes']['COD_S11'])

SEL1_set=set(df_venn.loc[df_venn.SEL==1]['COD_S11'])
SEL2_set=set(df_venn.loc[df_venn.SEL==2]['COD_S11'])
SEL3_set=set(df_venn.loc[df_venn.SEL==3]['COD_S11'])
SEL4_set=set(df_venn.loc[df_venn.SEL==4]['COD_S11'])
SEL_IHE1_set=set(df_venn.loc[df_venn.SEL_IHE==1]['COD_S11'])
SEL_IHE2_set=set(df_venn.loc[df_venn.SEL_IHE==2]['COD_S11'])
SEL_IHE3_set=set(df_venn.loc[df_venn.SEL_IHE==3]['COD_S11'])
SEL_IHE4_set=set(df_venn.loc[df_venn.SEL_IHE==4]['COD_S11'])

STRATUM_unknown_set=set(df_venn.loc[df_venn.STRATUM=='Unknown']['COD_S11'])
STRATUM_1_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 1']['COD_S11'])
STRATUM_2_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 2']['COD_S11'])
STRATUM_3_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 3']['COD_S11'])
STRATUM_4_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 4']['COD_S11'])
STRATUM_5_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 5']['COD_S11'])
STRATUM_6_set=set(df_venn.loc[df_venn.STRATUM=='Stratum 6']['COD_S11'])

```

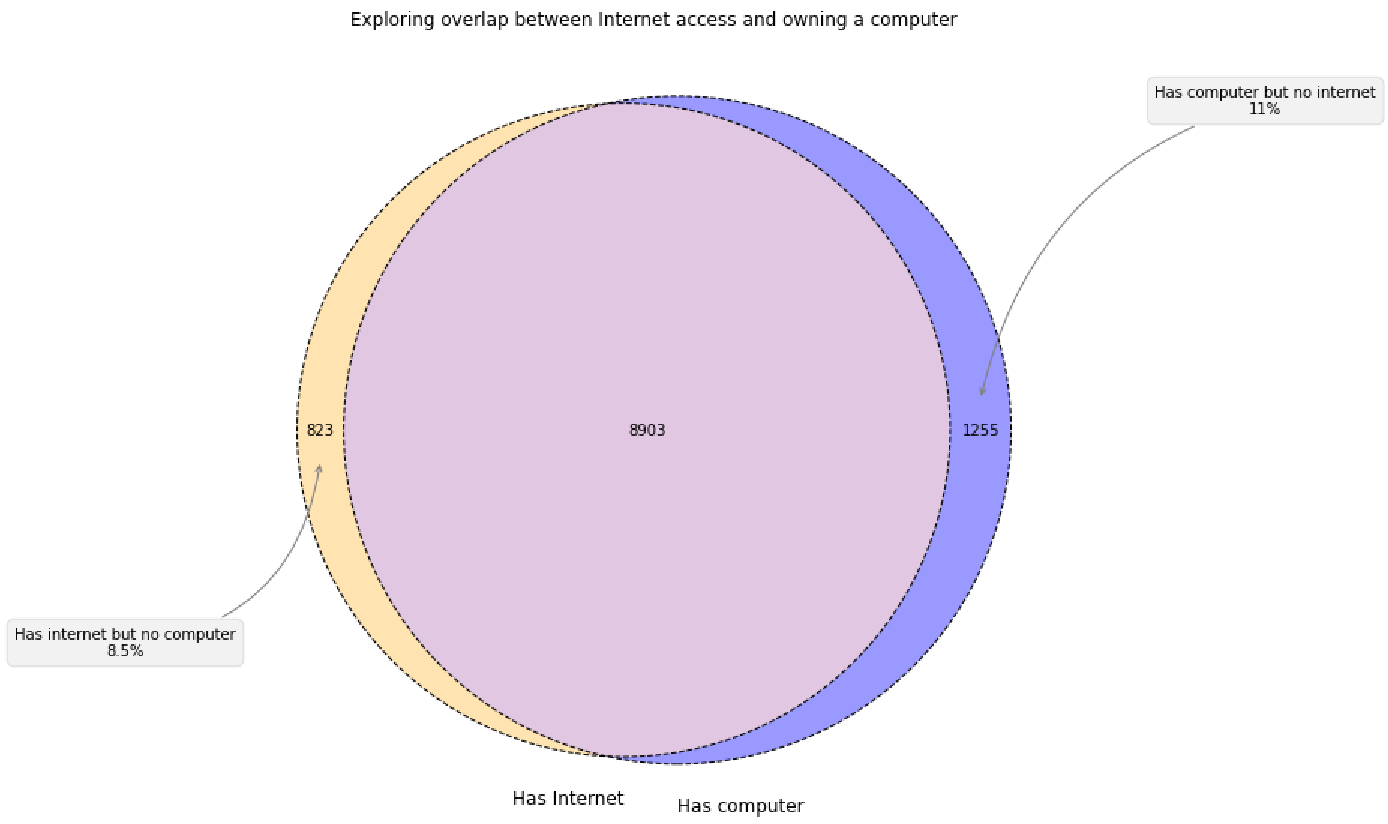
executed in 79ms, finished 18:20:09 2021-06-07

3 Venn diagrams, 2 sets

In [12]:

```
plt.figure(figsize=(10, 10))

sets=[Internet_set, Computer_set]
labels=('Has Internet', 'Has computer')

v=venn2([Internet_set, Computer_set], set_labels = labels, set_colors=("orange", "blue"))

v.get_patch_by_id('10').set_alpha(0.3)


venn2_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('Has internet but no computer\n8.5%',
             xy=v.get_label_by_id('10').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('Has computer but no internet\n11%',
             xy=v.get_label_by_id('01').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Exploring overlap between Internet access and owning a computer')
plt.show()
```

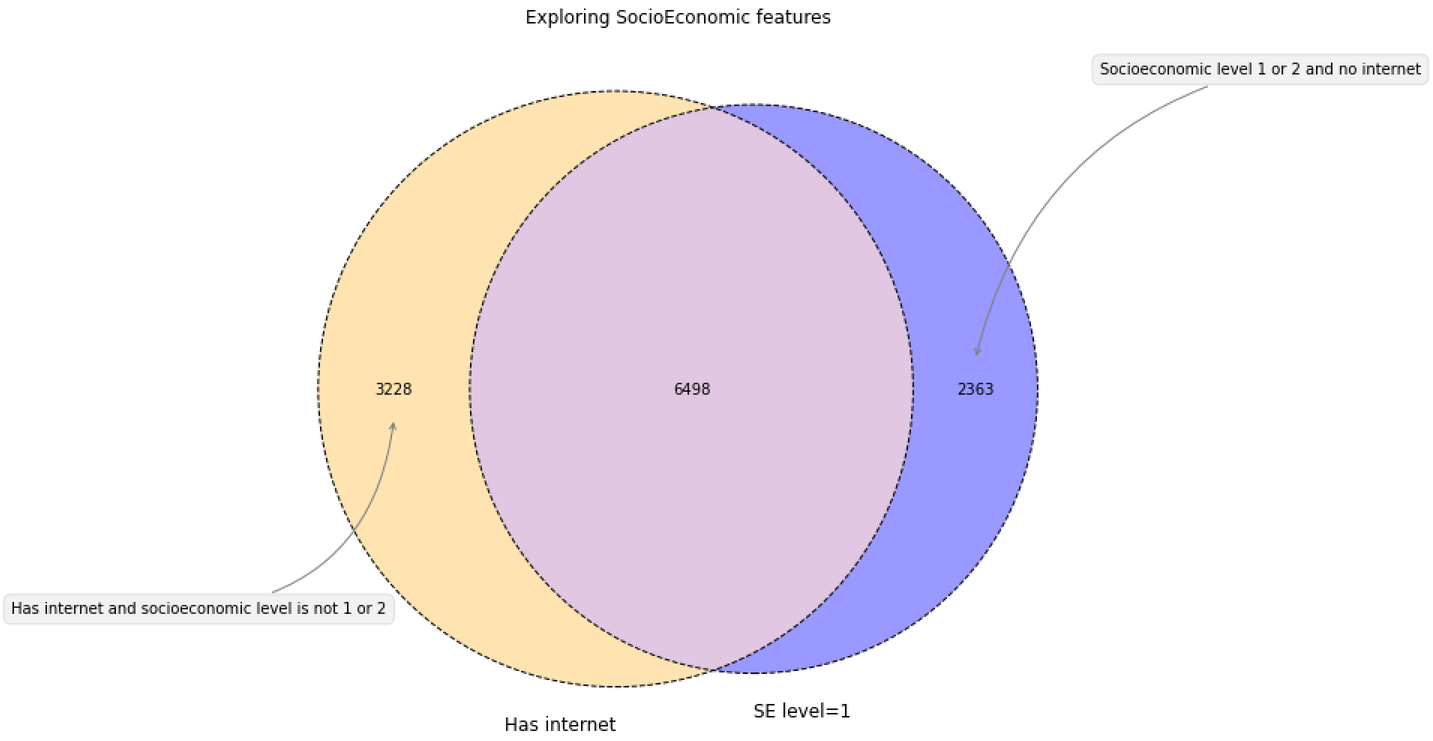
executed in 191ms, finished 18:20:09 2021-06-07



In [13]:

```python
plt.figure(figsize=(10, 10))

SEL1_and_SEL2_set=SEL_IHE1_set.union(SEL_IHE2_set)

sets=[Internet_set, SEL1_and_SEL2_set]
labels=('Has internet', 'SE level=1')

v=venn2(sets, set_labels = labels, set_colors=("orange", "blue"))

v.get_patch_by_id('10').set_alpha(0.3)


venn2_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('Has internet and socioeconomic level is not 1 or 2',
             xy=v.get_label_by_id('10').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('Socioeconomic level 1 or 2 and no internet',
             xy=v.get_label_by_id('01').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Exploring SocioEconomic features')
plt.show()
```

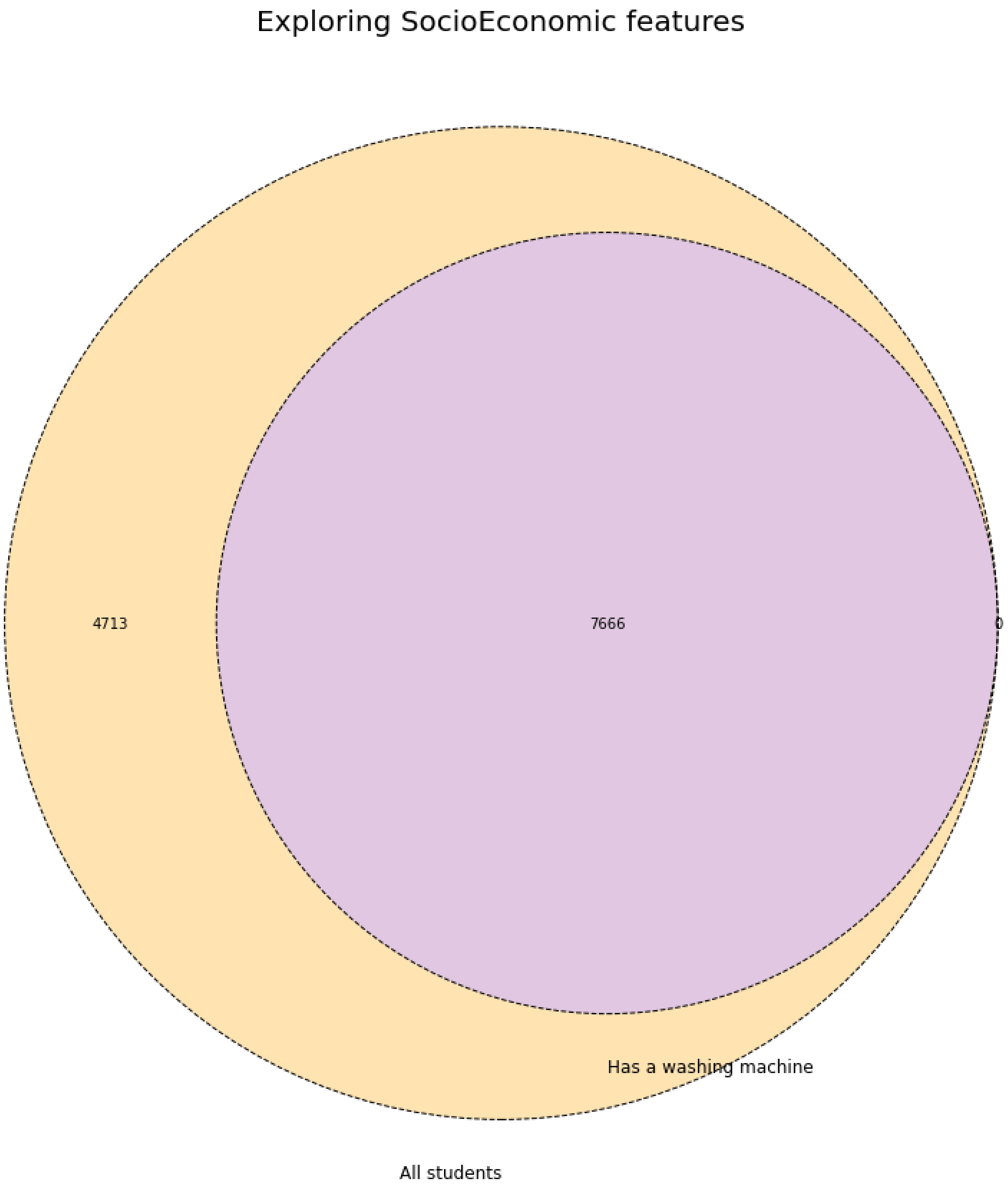
executed in 158ms, finished 18:20:09 2021-06-07



In [14]:

```
plt.figure(figsize=(15, 15))
ax=plt.gca()

sets=[ALL_set, Washing_machine_set]
labels=('All students', 'Has a washing machine')

v=venn2(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue"))

v.get_patch_by_id('10').set_alpha(0.3)


venn2_circles(subsets=sets,
              linestyle="dashed", linewidth=1)


plt.title('Exploring SocioEconomic features', fontsize='20')
plt.show()
```

executed in 158ms, finished 18:20:09 2021-06-07

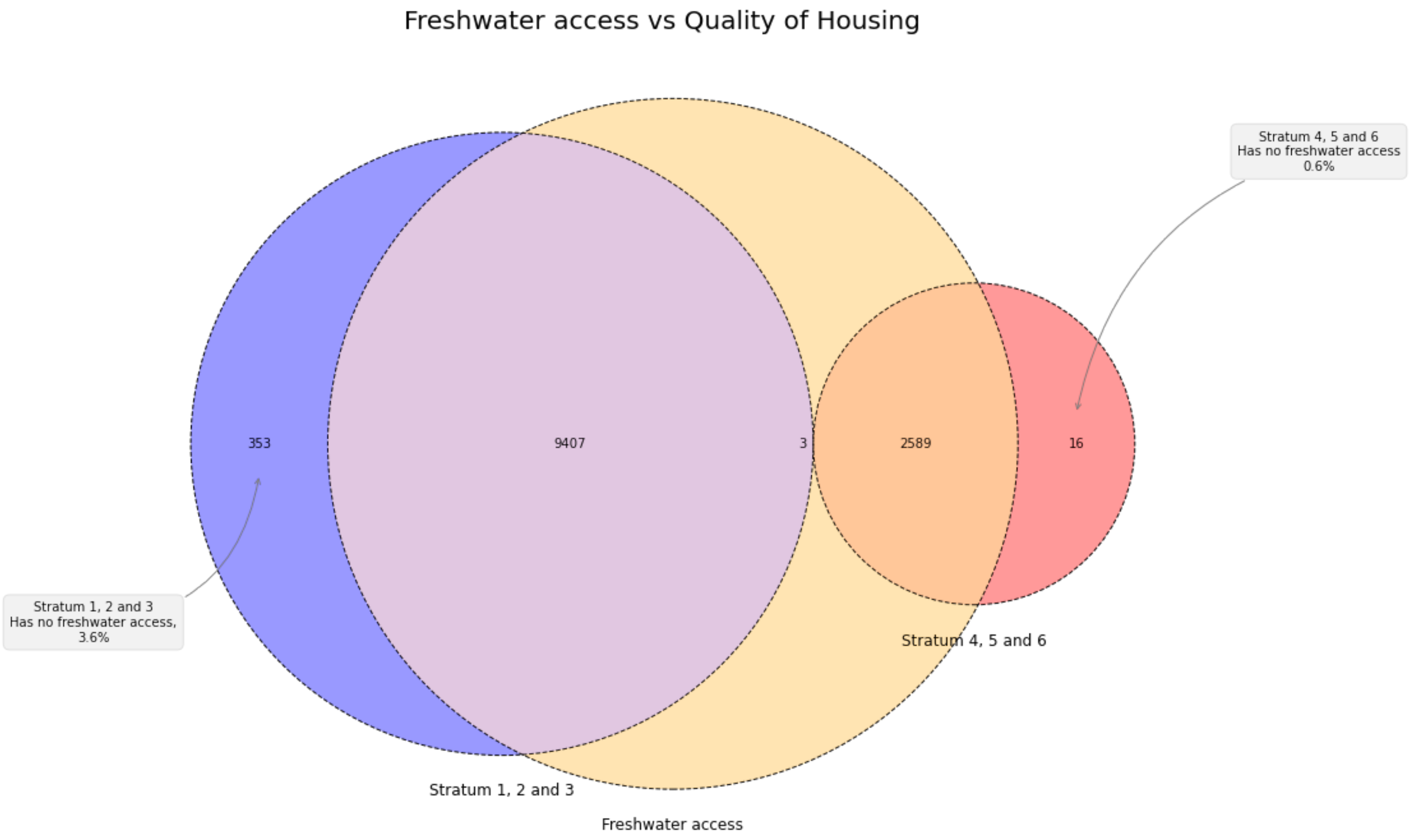


4 Venn diagrams, 3 sets

In [15]:

```
plt.figure(figsize=(15, 15))
ax=plt.gca()

Stratum1_Stratum2_Stratum3_set=STRATUM_1_set.union(STRATUM_2_set).union(STRATUM_3_set)
Stratum4_Stratum5_Stratum6_set=STRATUM_4_set.union(STRATUM_5_set).union(STRATUM_6_set)

sets=[Fresh_set, Stratum1_Stratum2_Stratum3_set, Stratum4_Stratum5_Stratum6_set]
labels=('Freshwater access', 'Stratum 1, 2 and 3', 'Stratum 4, 5 and 6')

v=venn3(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue", "red"))

v.get_patch_by_id('100').set_alpha(0.3)


venn3_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('Stratum 1, 2 and 3\nHas no freshwater access,\n3.6%',
             xy=v.get_label_by_id('010').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('Stratum 4, 5 and 6\nHas no freshwater access\n0.6%',
             xy=v.get_label_by_id('001').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Freshwater access vs Quality of Housing', fontsize='20')
plt.show()
```

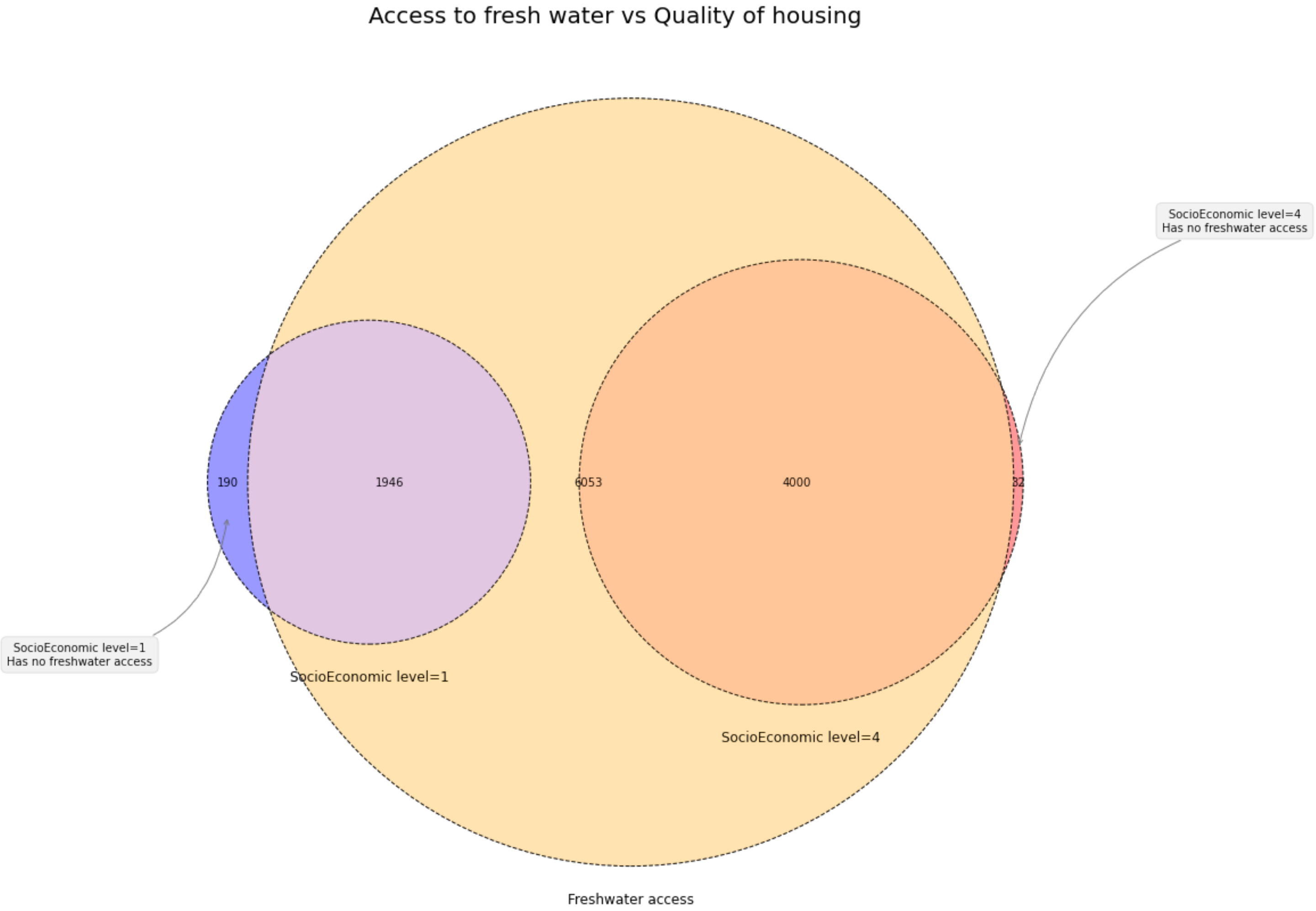
executed in 191ms, finished 18:20:09 2021-06-07



In [16]:

```python
plt.figure(figsize=(15, 15))
ax=plt.gca()



sets=[Fresh_set, SEL1_set, SEL4_set]
labels=('Freshwater access', 'SocioEconomic level=1', 'SocioEconomic level=4')

v=venn3(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue", "red"))

v.get_patch_by_id('100').set_alpha(0.3)


venn3_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('SocioEconomic level=1\nHas no freshwater access',
             xy=v.get_label_by_id('010').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('SocioEconomic level=4\nHas no freshwater access',
             xy=v.get_label_by_id('001').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Access to fresh water vs Quality of housing', fontsize='20')
plt.show()
```

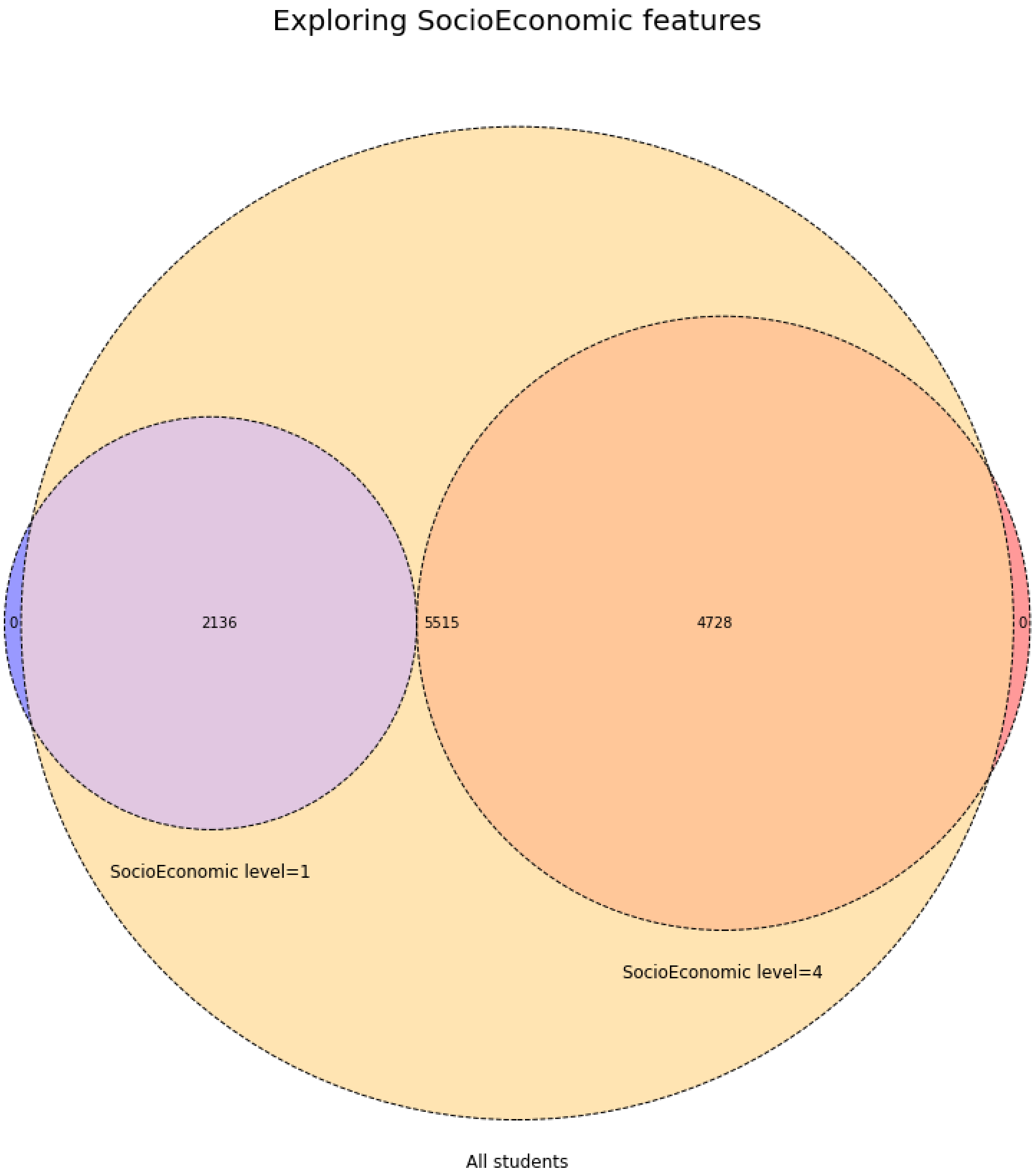
executed in 207ms, finished 18:20:10 2021-06-07



In [17]:

```python
plt.figure(figsize=(15, 15))
ax=plt.gca()

sets=[ALL_set, SEL1_set, SEL2_set]
labels=('All students', 'SocioEconomic level=1', 'SocioEconomic level=4')

v=venn3(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue", "red"))

v.get_patch_by_id('100').set_alpha(0.3)


venn3_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.title('Exploring SocioEconomic features', fontsize='20')
plt.show()
```

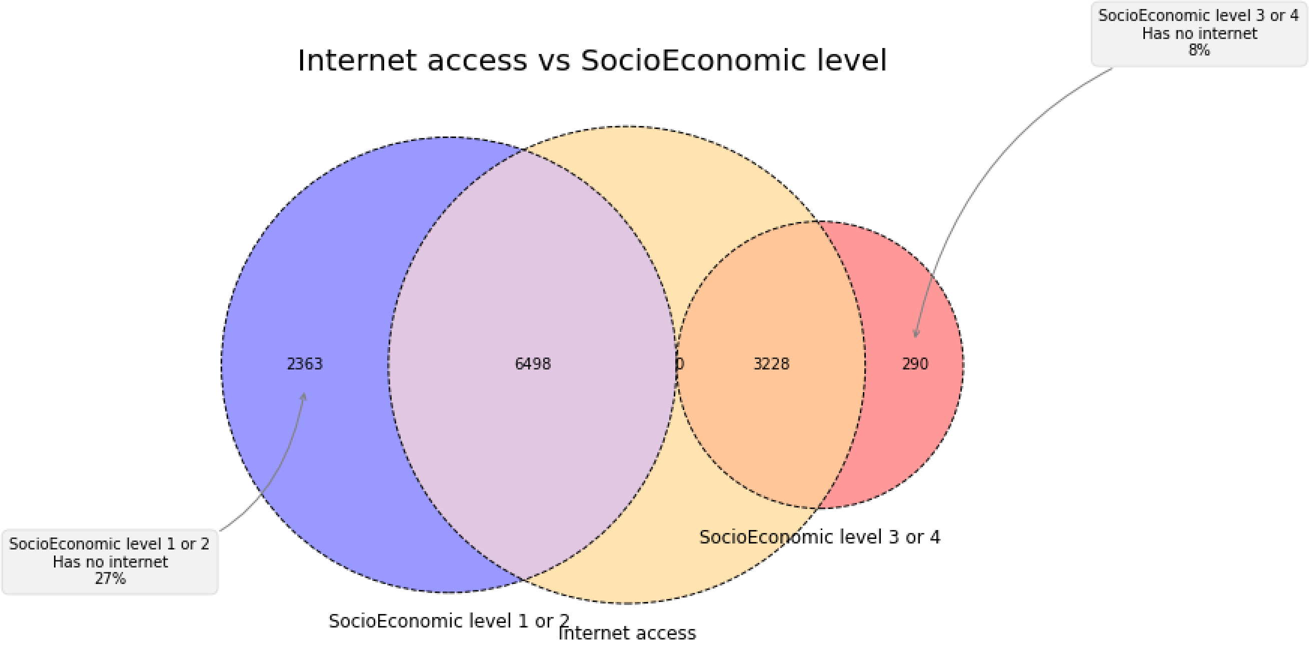
executed in 143ms, finished 18:20:10 2021-06-07



In [18]:

```
plt.figure(figsize=(10, 10))

SEL1_and_SEL2_set=SEL_IHE1_set.union(SEL_IHE2_set)
SEL3_and_SEL4_set=SEL_IHE3_set.union(SEL_IHE4_set)

ax=plt.gca()

sets=[Internet_set, SEL1_and_SEL2_set, SEL3_and_SEL4_set]
labels=('Internet access', 'SocioEconomic level 1 or 2', 'SocioEconomic level 3 or 4')

v=venn3(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue", "red"))

v.get_patch_by_id('100').set_alpha(0.3)


venn3_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('SocioEconomic level 1 or 2\nHas no internet\n27%',
             xy=v.get_label_by_id('010').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('SocioEconomic level 3 or 4\nHas no internet\n8%',
             xy=v.get_label_by_id('001').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Internet access vs SocioEconomic level', fontsize='20')
plt.show()
```

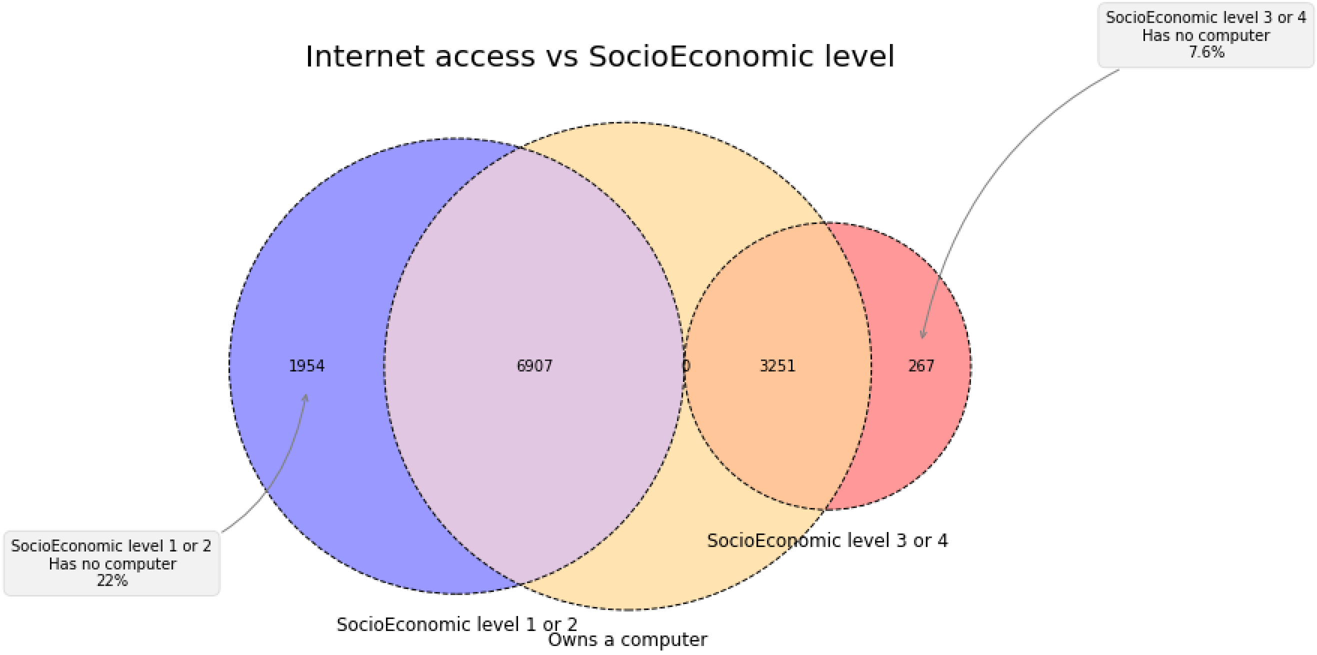
executed in 175ms, finished 18:20:10 2021-06-07



In [19]:

```
plt.figure(figsize=(10, 10))

ax=plt.gca()

sets=[Computer_set, SEL1_and_SEL2_set, SEL3_and_SEL4_set]
labels=('Owns a computer', 'SocioEconomic level 1 or 2', 'SocioEconomic level 3 or 4')

v=venn3(subsets=sets, set_labels = labels, ax=ax, set_colors=("orange", "blue", "red"))

v.get_patch_by_id('100').set_alpha(0.3)


venn3_circles(subsets=sets,
              linestyle="dashed", linewidth=1)

plt.annotate('SocioEconomic level 1 or 2\nHas no computer\n22%',
             xy=v.get_label_by_id('010').get_position() - np.array([0, 0.05]), xytext=(-130,-130),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.4',color='gray'))

plt.annotate('SocioEconomic level 3 or 4\nHas no computer\n7.6%',
             xy=v.get_label_by_id('001').get_position() - np.array([0, -0.05]), xytext=(190,190),
             ha='center', textcoords='offset points', bbox=dict(boxstyle='round, pad=0.5', fc='gray', alpha=0.1),
             arrowprops=dict(arrowstyle='->', connectionstyle='arc3,rad=0.3',color='gray'))


plt.title('Internet access vs SocioEconomic level', fontsize='20')
plt.show()
```

executed in 175ms, finished 18:20:10 2021-06-07



In [ ]:

```
```