Quiz 13 – **Solution** (Morning)

rdd1 = sc.parallelize([("david", 35), ("john", 25), ("mary", 45)])
rdd2 = sc.parallelize([("david", 5), ("david", 4), ("mary", 5), ("john", 4), ("mary", 3)])

1. [3 points] Find names of people whose ages are **less** than 30.

   rdd1.filter(lambda x: x[1] < 30).map(lambda x: x[0]).collect()

2. [3 points] Find names of people whose ages are **greater** than 30 and received a rating of 4 or above.

   R1 = rdd1.filter(lambda x: x[1] > 30)
   R2 = rdd2.filter(lambda x: x[1] >= 4)

   R1.join(R2).map(lambda x: x[0]).distinct().collect()

3. [4 points] Find, for each person, his/her **average** rating.

   rdd = rdd2.aggregateByKey((0,0), lambda U,v: (U[0] + v, U[1] + 1), lambda U1,U2: (U1[0] + U2[0], U1[1] + U2[1]))
   rdd.map(lambda (x, (y, z)): (x, float(y)/z)).collect()