CSE 562

Physical data organization
```python
from re import split

with open('data.csv', 'r') as f:
    for line in f:
        fields = split(', ', line)
        if fields[2] != "Ensign" and int(fields[3]) > 25:
            print(fields[1])
```

1, Redshirt, Ensign, 19,
2, Spock, Lt., 103,
3, Kirk, Capt., 22,
4, Redshirt, Ensign, 21,
5, Redshirt, Ensign, 18,
6, McCoy, Lt. Cmdr, 38,
f: sequence of records

(RAN) HDD

(HDD virtually)

HDD

page

page
Stream Iteration

Array

Stream of records

Stream of bytes

Pages
Idea 1

store fields in a native format

pro: No casting

con: Need to save information about storage format
Fixed size encoding

Pros: No delimiters (saves 4 bytes)

Cons: Need to know max length

Wasted space on variable length fields
Not good for adding data
Representation records

"CSV" -> delimiters

"Fixed Width"

"Per record directory"

Factors

Size of the file

Types of the fields

How much data is changing
read file
filter records
transform
output