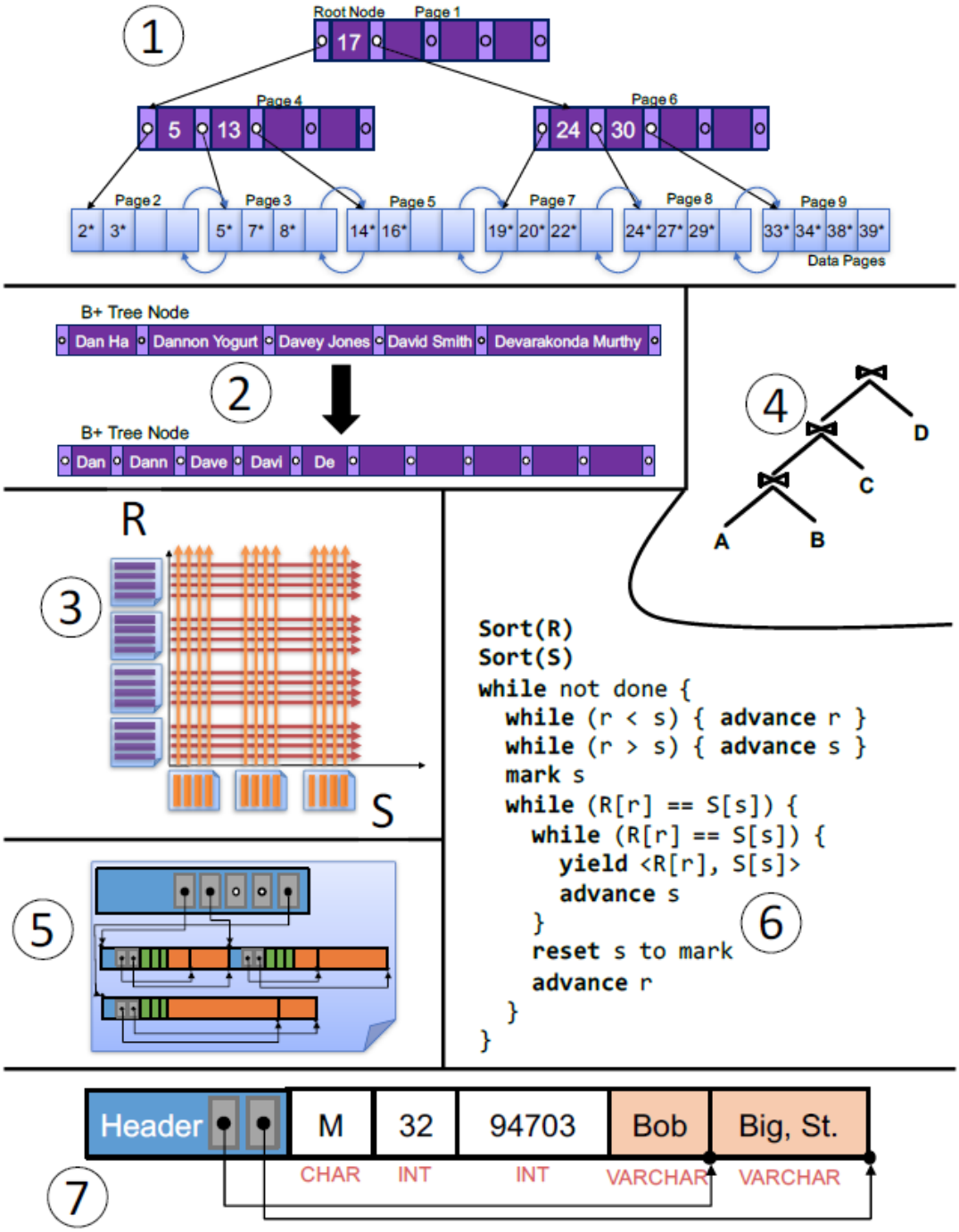
I Basics [10 points]

For each image on the next page, select the letter corresponding to the best description.
A. Left Deep Tree
B. Key Compression
C. B+ Tree
D. ISAM
E. Nested Loops Join
F. Sort Merge Join
G. Indexed Nested Loop Join
H. Slotted Page
I. Variable Length Tuple
J. Fixed Length Tuple
K. Buffer Frame
L. Sort based group by
M. mapPartitions
N. Tournament Sort


1
Root Node
Page 1
17
Page 4
5
13
Page 6
24
30
Page 2
2*
3*
Page 3
5*
7*
8*
Page 5
14*
16*
Page 7
19*
20*
22*
Page 8
24*
27*
29*
Page 9
33*
34*
38*
39*
Data Pages
B+ Tree Node
Dan Ha
Dannon Yogurt
Davey Jones
David Smith
Devarakonda Murthy
2
B+ Tree Node
Dan
Dann
Dave
Davi
De
4
D
C
A
B
R
3
S
Sort(R)
Sort(S)
while not done {
while (r < s) { advance r }
while (r > s) { advance s }
mark s
while (R[r] == S[s]) {
while (R[r] == S[s]) {
yield <R[r], S[s]>
advance s
}
reset s to mark
advance r
}
}
5
6
Header
M
32
94703
Bob
Big, St.
CHAR
INT
INT
VARCHAR
VARCHAR
7

II SQL and ER Modeling [12 points]

1. **[2 points]** Which of the following expressions computes the matrix vector product:

$$(\mathbf{Ax})_i = \sum_{k=1}^{d} A_{ik} x_k$$

Assume **A** and **x** have compatible dimensions and there is only one correct answer.
A.

```
SELECT A.row AS row, A.value * x.value AS value
FROM A JOIN x
ON A.row = x.row
```

B.

```
SELECT A.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.row = x.row
GROUP BY A.col
```

C.

```
SELECT x.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.col = x.row
GROUP BY A.col
```

D.

```
SELECT A.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.col = x.row
GROUP BY A.row
```

2. **[2 points]** Suppose we wanted to compute the element-wise sum of the vectors **x** and **y** using the SQL expression:

```
SELECT x.row AS row, SUM(x.value + y.value) AS value
FROM x JOIN y
ON x.row = y.row
```

Which of the following statements about this query are true? (You may mark zero (ϕ), one or more than one of the choices.)
A. Some non-zero entries may be omitted from the final result.
B. The correct query should use LEFT OUTER JOIN.
C. The correct query should use FULL OUTER JOIN.
D. There is nothing wrong.

3. **[8 points]** There are four tables. SALESPERSON contains the names, ids & quotas for the salespeople, and names are not unique. PRODUCTS contains the product names, product ids, and prices for the products. The product ids are unique. CUSTOMERS contains the customer names, customer ids, and regions for the customers (customer ids are unique), and ORDERS contains the customer id, the product id, and the product ordered per customer.

SALESPERSON

Sname	Sid	Quota
Frances	25	$100
Bob	31	$150
Frances	74	$200
Mary	89	$250

PRODUCTS

Pname	Pid	Pprice
disks	131	$100
pcs	152	$700
macs	831	$800
printers	255	$120
paper	221	$5

CUSTOMERS

Cname	Cid	Region
Bob	1	TX
Harry	2	TX
Lin	3	MA
Martha	4	FL
Lin	5	FL
Leyla	6	CA

ORDERS

Cid	Pid	Quantity
1	152	1
2	152	1
4	831	1
4	131	1
5	255	1
6	831	1

(a) **[2 points]** Select the true SQL expression(s) for "List the names of the customers who have bought more than one item." (You may mark zero (ϕ), one or more than one of the choices.)

A.

```
SELECT cname
FROM customers
WHERE cid IN (SELECT cid
              FROM orders
              GROUP BY cid
              HAVING count(*) > 1)
```

B.

```
SELECT c.cname
FROM customers c, (SELECT cid
                   FROM orders
                   GROUP BY cid
                   HAVING count(*) > 1) as o
WHERE c.cid = o.cid
```

C.

```
SELECT cname
FROM customers
WHERE (SELECT count(*)
       FROM orders
       GROUP BY pid ) > 1
```

D.

```
SELECT cname
FROM customers c, orders o1, orders o2
WHERE c.cid = o1.cid and c.cid = o2.cid and o1.cid <> o2.pid
```

(b) **[2 points]** Select the true SQL expressions for "List the names, pid, and price of all the products, whether or not the product has been ordered, but if it has been ordered by the cids of the customer who ordered it." (You may mark zero (ϕ), one or more than one of the choices.)

A.

```
SELECT name, pid, price, cid
FROM products LEFT OUTER JOIN orders
```

B.

```
SELECT name, pid, price, cid
FROM products LEFT OUTER JOIN orders
ON products.pid = orders.pid
```

C.

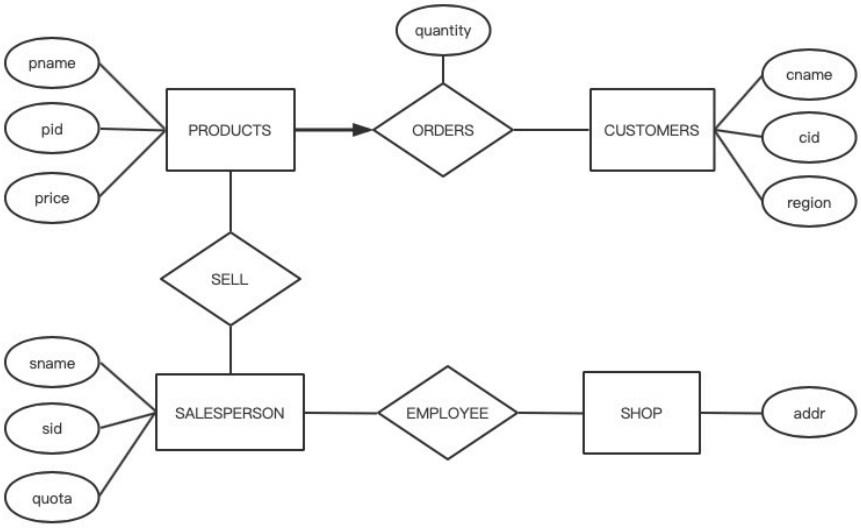
```
SELECT name, p.pid, price, cid
FROM products p, orders o
WHERE p.pid = o.pid
UNION
SELECT name, pid, price
FROM products p
WHERE NOT EXISTS (SELECT * FROM orders o WHERE o.pid = p.pid)
```

D.

```
SELECT name, p.pid, price, cid
FROM products p, orders o
WHERE p.pid = o.pid
UNION
SELECT name, pid, price, NULL as cid
FROM products p
```

(c) **[4 points]** Now, let's complete the ER diagram with the tables given above and some new translations. You need to underline the primary keys, add arrows. If bolding a line/arrow, be sure to clearly make it bold. If here's weak entity, also be sure to clearly show it.

- Each product sold by exactly one salesperson.
- The sid of salesperson in the same shop is unique, however two salesperson in different shop might have the same sid.
- A salesperson can only work in one shop.
- The address can uniquely identify a shop.



III Indexes and B+ Trees [9 points]

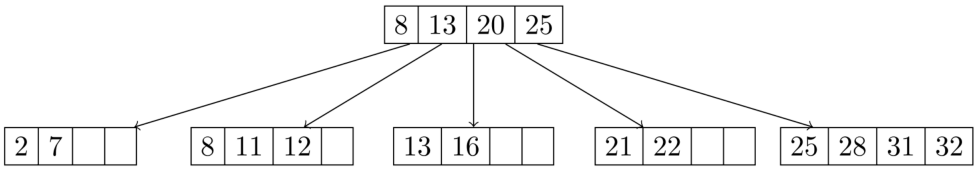
1. **[5 points]** Basics of B+ Trees.

   (a) **[1 point]** What is the maximum fan out F of an order d B+ tree?

   (b) **[2 points]** Assume that each leaf can hold F data entries. What is the maximum number of data entries that an order d B+ tree of height 5 can store? Express your answer as a function of F. (Note that a height 1 B+ tree only has a root node).

   (c) **[2 points]** Again, assume that each leaf can hold F data entries. And we have indexed a file with 1×10^9 records with this order d B+ tree. What is the minimum number of I/Os it will take to check if a data entry exists in this tree? Express your answer as a function of F.

2. **[4 points]** Consider the following B+ tree of order 2.



   (a) **[1 point]** How many nodes split when you insert 27?

   (b) **[1 point]** After inserting 27 into the tree, you also insert 26. How many nodes split as a result of inserting 26?

   (c) **[2 points]** Assume that after inserting 26 and 27, you insert the keys 34, 35, 36, . . ., 100. After all these insertions, what keys are in the leftmost leaf node?

IV External Sorting [8 points]

1. **[2 points]** True or False:

   (a) Increasing the number of buffer pages don't affects the number of I/Os performed in Pass 0 of an external sort.

   (b) Double buffering reduces the time it takes to sort records within a single page.

2. **[6 points]** Assume that each page is 4 KB large, and that you have a 24KB buffer pool (with 6 frames).

   (a) **[2 points]** How many passes would it take to externally sort an 512KB file? Include the initial sorting pass and subsequent merging passes in your answer. You need to simplify your answer.

   (b) **[2 points]** What would be the total cost in I/Os for this external sort?

   (c) **[2 points]** What is the minimum number of additional buffer frames we require to reduce the number of passes (from part 1) by 1?

V Join Algorithms [8 points]

1. **[5 points]** Consider a relation R with attributes (x, y) and a relation S with attributes (y, z). Column y in S is a key and the set of values of y in R are the same as the set of values of y in S. Assume that there are no indexes available and that there are 25 pages in the buffer available. Table R is 1,500 pages with 50 tuples per page. Table S is 400 with 100 tuples per page. Compute the I/O costs for the following joins. Assume the simplest cost model, where pages are read and written one at a time.

   (a) **[1 point]** Block nested loops join with R as the outer relation and S as the inner relation.

   (b) **[1 point]** Block nested loops join with S as the outer relation and R as the inner relation.

   (c) **[1 point]** Sort merge join with R as the outer relation and S as the inner relation.

   (d) **[1 point]** Sort merge join with S as the outer relation and R as the inner relation.

   (e) **[1 point]** Hash join with S as the outer relation and R as the inner relation.

2. **[3 points]** Consider a new case, i.e. B>4 pages worth of buffer space, and relations M and N of size > B. Please fill the blanks below with "always", "sometimes" or "never".

   (a) **[1 point]** Block nested loop join is _____ better than page-oriented nested loop join.

   (b) **[1 point]** Sort-merge join is _____ better than hash-join.

   (c) **[1 point]** Hybrid Hash-Join is _____ better than block-nested loops join.

VI Query Optimization [10 points]

Consider two relations Cat(age, weight, price) and Pocket(money), with 150 tuples and 100 tuples respectively. We have an index on Cat.age with 15 unique integer values uniformly distributed in the range [1, 15], an index on Cat.weight with 30 unique float values uniformly distributed in the range [2001, 5000], an index on Cat.price with 10 unique integer values uniformly distributed in the range [11, 20], and an index on Pocket.money with 15 unique integer values uniformly distributed in the range [11, 25].
Use selectivity estimation to estimate the number of tuples produced by the following queries.

(a) **[1 point]** SELECT * FROM Cat

(b) **[2 points]** SELECT * FROM Cat WHERE age $\geq$ 10

(c) **[2 points]** SELECT * FROM Cat WHERE age $<$ 5 AND weight $\leq$ 3000

(d) **[2 points]** SELECT * FROM Cat WHERE age $>$ 10 OR price $\geq$ 15

(e) **[3 points]** SELECT * FROM Cat, Pocket WHERE Cat.price = Pocket.money

VII Transaction and Concurrency [10 points]

Consider the following schedule. (For each of the questions below, you may mark zero (ϕ), one or more than one of the choices.)

	T1	T2	T3	T4
1	R(A)			
2		R(A)		
3			R(C)	
4			W(C)	
5		R(B)		
6		W(B)		
7	R(B)			
8				R(B)
9	R(C)			
10				R(C)
11				W(B)
12		commit		
13			commit	
14	commit			
15				commit

(a) **[1 point]** What transactions is T1 pointing to in the conflict graph for the schedule above?
A. T1
B. T2
C. T3
D. T4

(b) **[1 point]** What transactions is T2 pointing to in the conflict graph for the schedule above?
A. T1
B. T2
C. T3
D. T4

(c) **[1 point]** What transactions is T3 pointing to in the conflict graph for the schedule above?
A. T1
B. T2
C. T3
D. T4

(d) **[1 point]** What transactions is T4 pointing to in the conflict graph for the schedule above?
A. T1
B. T2
C. T3
D. T4

(e) **[3 points]** Which of the following locking disciplines could have produced the above schedule?
A. 2 phase locking
B. Strict 2 phase locking

(f) **[3 points]** Which of the following schedules below are conflict equivalent to the schedule above?
A. T3, T1, T2, T4
B. T2, T3, T1, T4
C. T4, T3, T1, T2
D. T1, T2, T3, T4
E. T3, T2, T1, T4

VIII Logging and Recovery [13 points]

1. **[5 points] General Logging and Recovery.**
   Mark the boxes for all true statement(s):

   (a) Schedules produced by two phase locking are guaranteed to prevent cascading aborts.

   (b) Strict two phase locking is both necessary and sufficient to guarantee conflict serializability.

   (c) In a system that uses strict two-phase locking, if a transaction aborts, it releases all of its locks as soon as rollback is complete.

   (d) In a system that uses strict two-phase locking, a transaction that only performs reads can never enter a deadlock cycle.

   (e) When aborting a transaction, it is necessary to modify pages on disk.

   (f) During recovery, the ARIES protocol redo aborted transactions.

   (g) When a transaction commits, any modified buffer pages must be written to durable storage.

   (h) In ARIES recovery, after the analysis phase, the recLSN of each page in the dirty page table must be larger than the pageLSN of the corresponding page.

   (i) If PageLSN is greater than the max LSN flushed so far (flushedLSN), we can safely write this page to disk.

   (j) Write-Ahead Logging (WAL) guarantees that a transactions log records are flushed to disk before the transaction commit.

2. **[8 points] Recovery.**
   Your database server has just crashed due to a power outage. You boot it back up, find the following log and checkpoint information on disk, and begin the recovery process. Assume we use a STEAL/NO FORCE recovery policy.

LSN	Record	prevLSN
30	update: T3 writes P5	null
40	update: T4 writes P1	null
50	update: T4 writes P5	40
60	update: T2 writes P5	null
70	update: T1 writes P2	null
80	Begin Checkpoint	-
90	update: T1 writes P3	70
100	End Checkpoint	-
110	update: T2 writes P3	60
120	T2 commit	110
130	update: T4 writes P1	50
140	T2 end	120
150	T4 abort	130
160	update: T5 writes P2	Null
180	CLR: undo T4 LSN 130	150

Transaction table at time of checkpoint

Transaction ID	lastLSN	Status
T1	70	Running
T2	60	Running
T3	30	Running
T4	50	Running

Dirty page table at time of checkpoint

Page ID	recLSN
P5	50
P1	40

   (a) **[2 points]** At the end of the Analysis phase, what transactions will be in the transaction table, and with what lastLSN and Status values?

Transaction ID	lastLSN	status

Table 1: Transaction table

   (b) **[2 points]** At the end of the Analysis phase, what pages will be in the dirty page table, and with what recLSN values?

page ID	recLSNs

Table 2: Dirty page table

(c) **[4 points]** At which LSN in the log should redo begin? Which log records will be redone (list their LSNs)?

IX Data Mining and Machine Learning [20 points]

1. **[4 points] General Data Mining and Machine Learning.**
   Mark the boxes for all true statement(s):

   (a) In multi-dimensional data model, it contains one fact table and multiple dimension tables.

   (b) The order of the basic KDD (Knowledge Discovery in Database) process is Data Selection, Data Cleaning, Evaluation, Data Mining & ML.

   (c) In supervised machine learning labeled data is used to train a model.

   (d) Classification models would be a good candidate when trying to predict the total amount of time a user spends browsing a page.

   (e) The ultimate goal in machine learning is to find a model that best fits the training data.

   (f) The k-means algorithm is guaranteed to converge to the global optimum.

   (g) The bag-of-words model encodes text as a vector.

   (h) A common form of feature engineering on continuous data is one-hot-encoding.

2. **[8 points] K-Means.**

   (a) **[2 points]** True or False:
   Suppose we are going to cluster the following dataset: $\{(x_1, y_1), ..., (x_n, y_n)\}$, denote $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, we can have a decentralized dataset $\{(x'_1, y'_1), ..., (x'_n, y'_n)\}$ where $x'_i = x_i - \bar{x}$, $y'_i = y_i - \bar{y}$, $i = 1, 2, ..., n$. The k-means algorithm will converge to the same result if we choose the initial centers as $\{(x_s, y_s), (x_t, y_t)\}$ for the original dataset and $\{(x'_s, y'_s), (x'_t, y'_t)\}$ for the decentralized dataset, where s, t are two different constants.

   (b) **[2 points]** Which of the following can act as possible termination conditions in k-means? (you may mark zero (ϕ), one or more than one of the choices.)
   A. For a fixed number of iterations.
   B. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
   C. Centroids do not change between successive iterations.
   D. Terminate when RSS falls below a threshold.

   (c) **[2 points]** In which of the following cases will k-means clustering fail to give good results? (you may mark zero (ϕ), one or more than one of the choices.)
   A. Data points with outliers.
   B. Data points with different densities.
   C. Data points with round shapes.
   D. Data points with non-convex shapes.

   (d) **[2 points]** The following is a set of one-dimensional points: $\{-4, 0, 1, 2, 3, 5, 7, 10, 13, 22\}$, perform two iterations of k-means on these points using the two initial centroids 1 and 7.
   1) What are the two new centroids after the first iteration?
   2) What are the two new centroids after the second iteration?

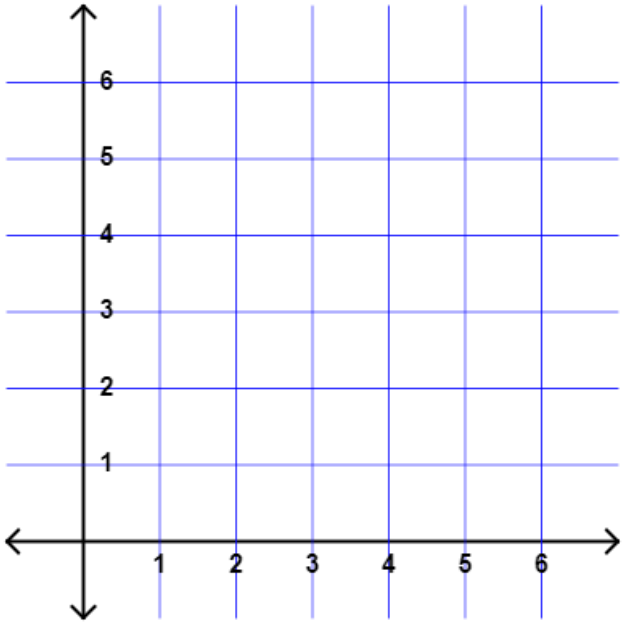
3. **[8 points] Linear Regression.**
   Given a set of i.i.d. data points $(x_1, y_1) \cdots (x_n, y_n)$, where $x \in \mathbb{R}$ and $y \in \mathbb{R}$ denote the input feature and the output response, respectively. By assuming the linear model is a reasonable approximation, we consider fitting the model via least squares approaches, in which we choose coefficients θ and θ_0 to minimize the Residual Sum of Squares (RSS),

$$\hat{\theta},\ \hat{\theta}_0 = \underset{\theta,\ \theta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \theta x_i - \theta_0)^2. \tag{1}$$

   Based on 5 observations of (x, y):

$$(1, 2), \quad (2, 3), \quad (3, 5), \quad (4, 4), \quad (5, 6), \tag{2}$$

   please answer the following questions:



(a) **[1 point]** Draw a 6 by 6 space with all the 5 data points.

(b) **[3 points]** Estimate the model parameters θ and θ_0.

(c) **[2 points]** Using (1), argue that the least squares line,

$$y = \hat{\theta}x + \hat{\theta}_0, \tag{3}$$

always passes through the points $(0, \hat{\theta}_0)$ and $(\bar{x}, \bar{y})$, where $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i$ and $\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i$, and plot the line in the picture of (a).

(d) **[2 points]** Find the data point which makes the highest contribution to RSS, and show the geometric interpretation in the picture of (a).