Hermit: Designing Succinct Secondary Indexing Mechanism by Exploiting Column Correlations

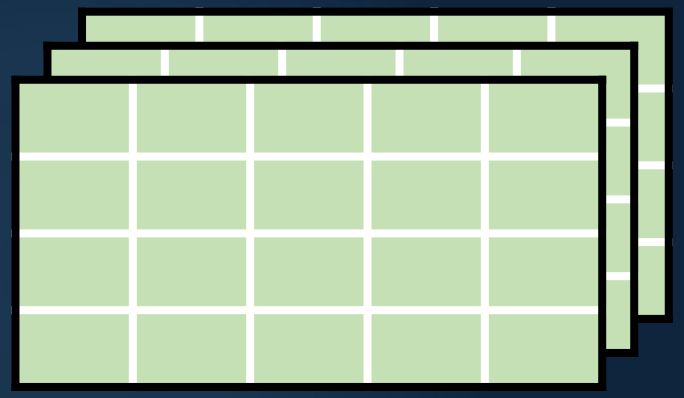
Yingjun Wu, Jia Yu, Yuanyuan Tian, Richard Sidle, Ronald Barber

IBM Almaden Research Center

Arizona State University

HERMIT: MOTIVATION

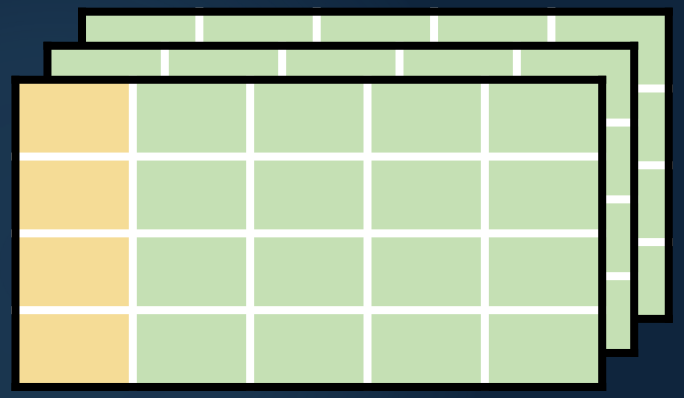
- An index is a copy of selected columns of data from a table



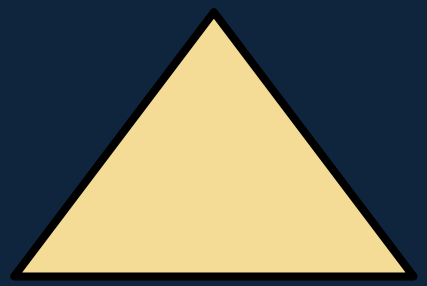
Base table

HERMIT: MOTIVATION

- An index is a copy of selected columns of data from a table



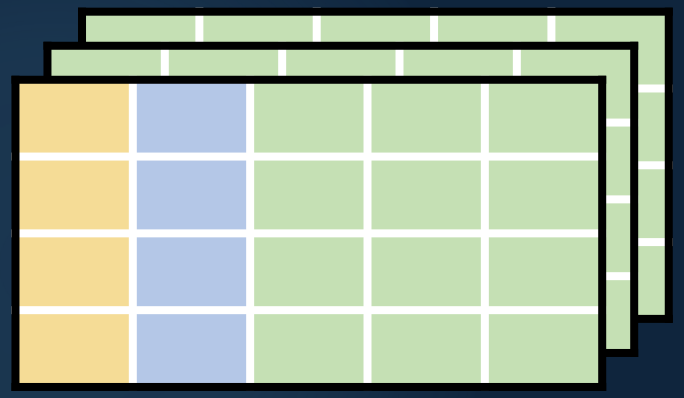
Base table



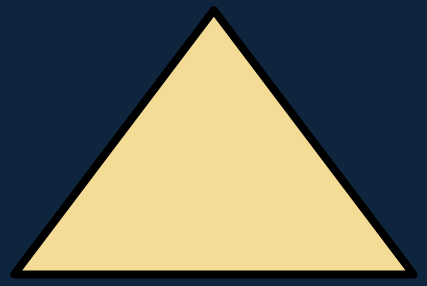
Primary index

HERMIT: MOTIVATION

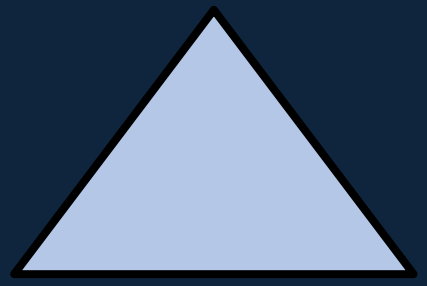
- An index is a copy of selected columns of data from a table



Base table



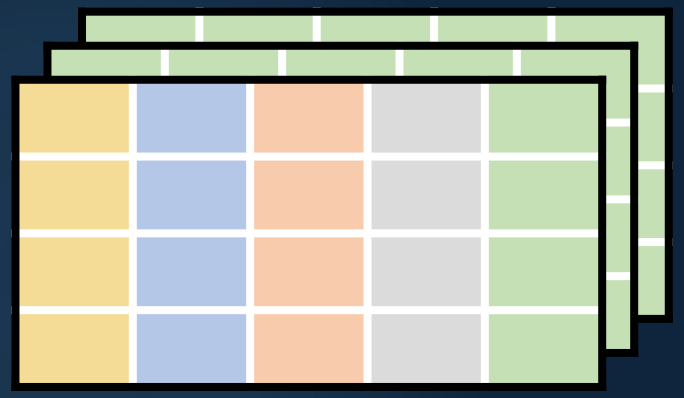
Primary index



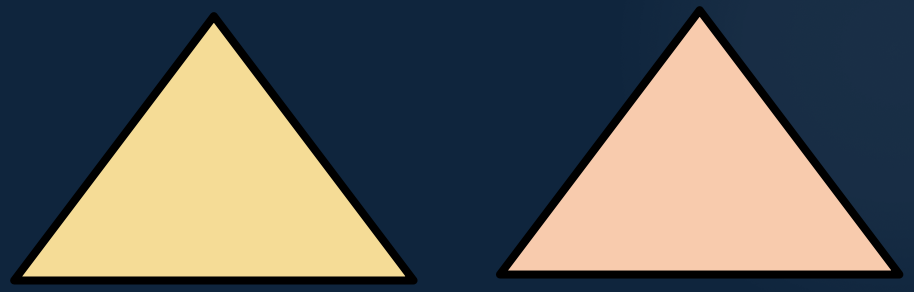
Secondary index

HERMIT: MOTIVATION

- An index is a copy of selected columns of data from a table



Base table



Primary index

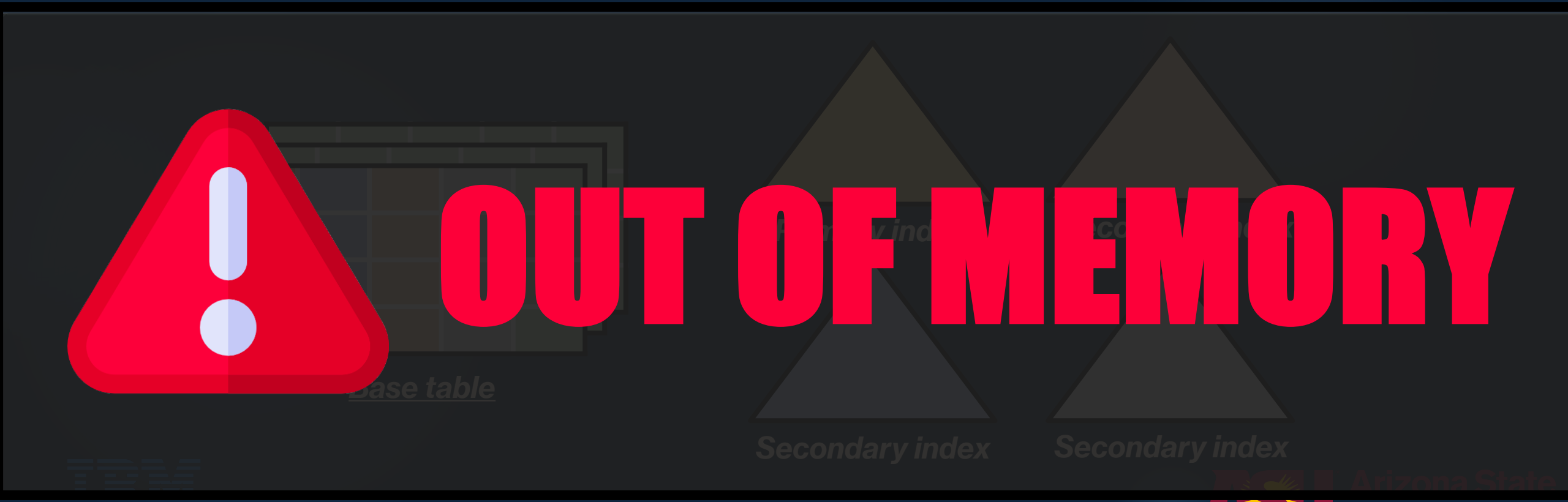
Secondary index

Secondary index

Secondary index

HERMIT: MOTIVATION

- An index is a copy of selected columns of data from a table



HERMIT: MOTIVATION

Indexes consume large amounts of space

B+ Tree on TPC-H	Index size	% space
2 GB	0.25 GB	
20 GB	2.51 GB	**12.5%**
200 GB	25 GB	

Source: ***Jia Yu and Mohamed Sarwat***. *"Two Birds, One Stone: A Fast, yet Lightweight, Indexing Scheme for Modern Database Systems" VLDB 2016.*

HERMIT: MOTIVATION

- **Existing solutions**
  - Sparse index, index compression…
    - Reduce space usage
    - At the expense of reduced lookup performance
    - Previous work Hippo index (VLDB 2016)
      - 50X smaller than B-Tree
      - Only comparable to B-Tree at 0.1% selectivity
  - Smart index selection
    - Low performance on unindexed columns

Disk Page Range	Summary
1 - 10	Synopsis, histogram…
11 – 25	Synopsis, histogram…
26 - 30	Synopsis, histogram…

Source: ***Jia Yu and Mohamed Sarwat***. *"Two Birds, One Stone: A Fast, yet Lightweight, Indexing Scheme for Modern Database Systems" VLDB 2016.*

HERMIT: OBJECTIVE

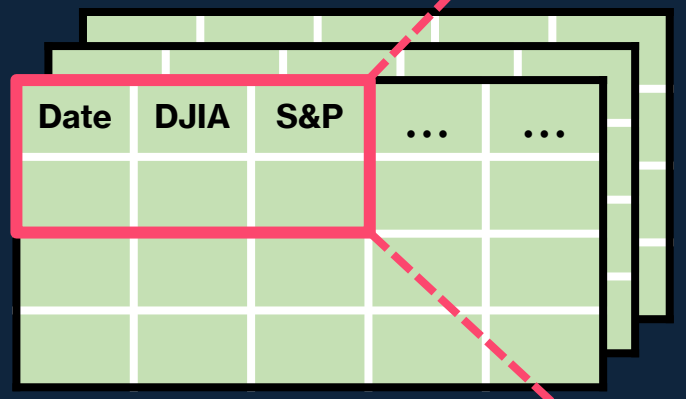
Consume much less space
Achieve good enough performance

HERMIT: OBSERVATION

Column correlations **are prevalent**

HERMIT: OBSERVATION

Column correlations **are prevalent**



STOCK table

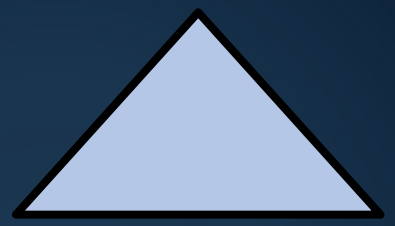
DJIA: Dow Jones Industrial Average

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

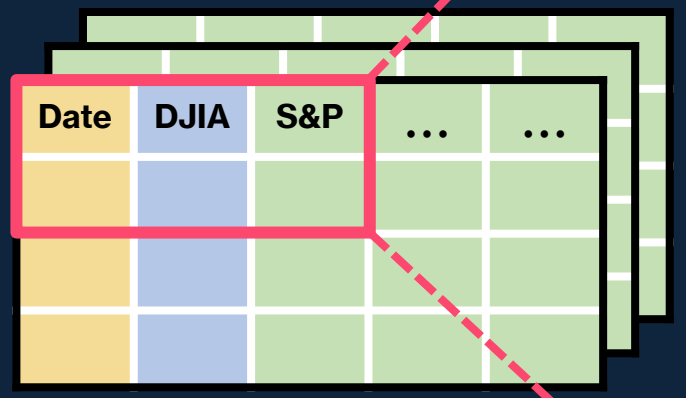
HERMIT: OBSERVATION

Column correlations are prevalent

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)



Secondary index for Date+DJIA



STOCK table

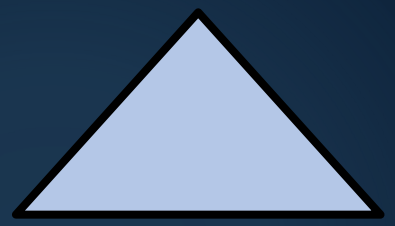
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: OBSERVATION

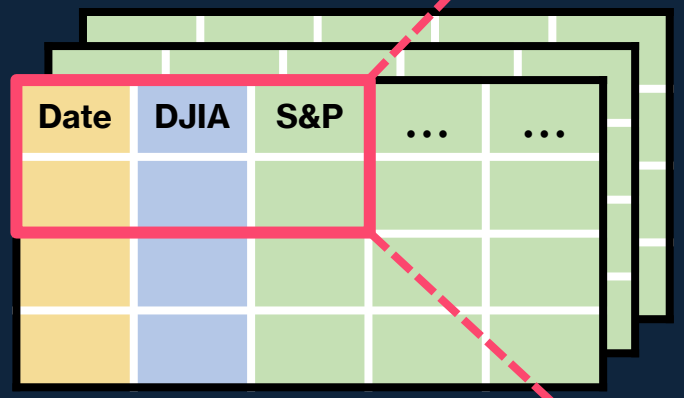
Column correlations are prevalent

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA



STOCK *table*

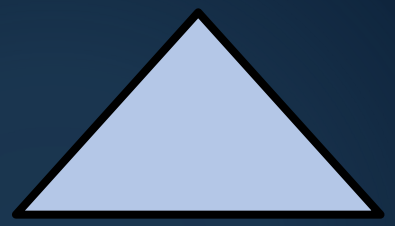
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: OBSERVATION

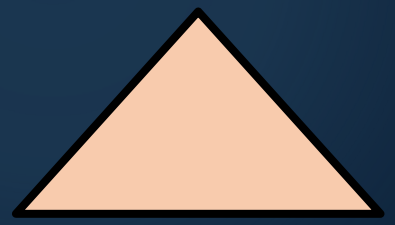
Column correlations are prevalent

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

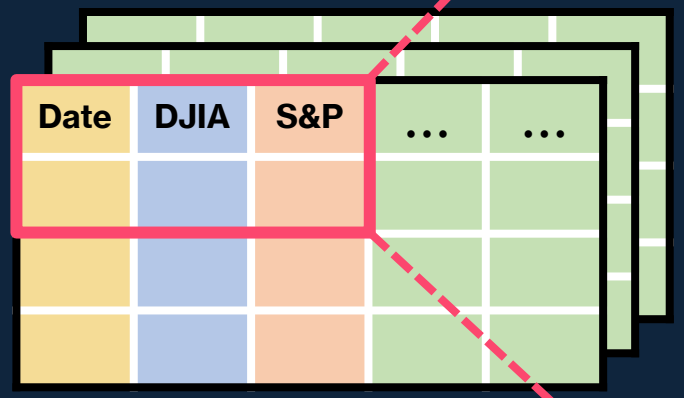
QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA



Secondary index for Date+S&P



STOCK table

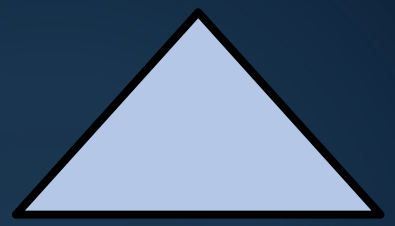
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: OBSERVATION

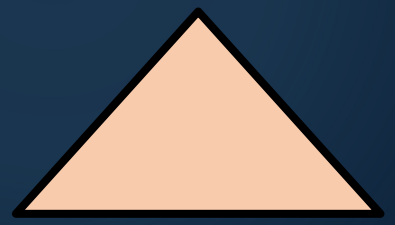
Column correlations **are prevalent**

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

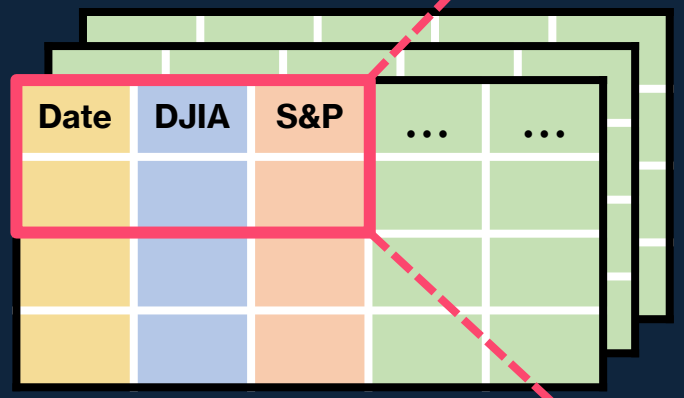
QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA



Secondary index for Date+S&P



STOCK table

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

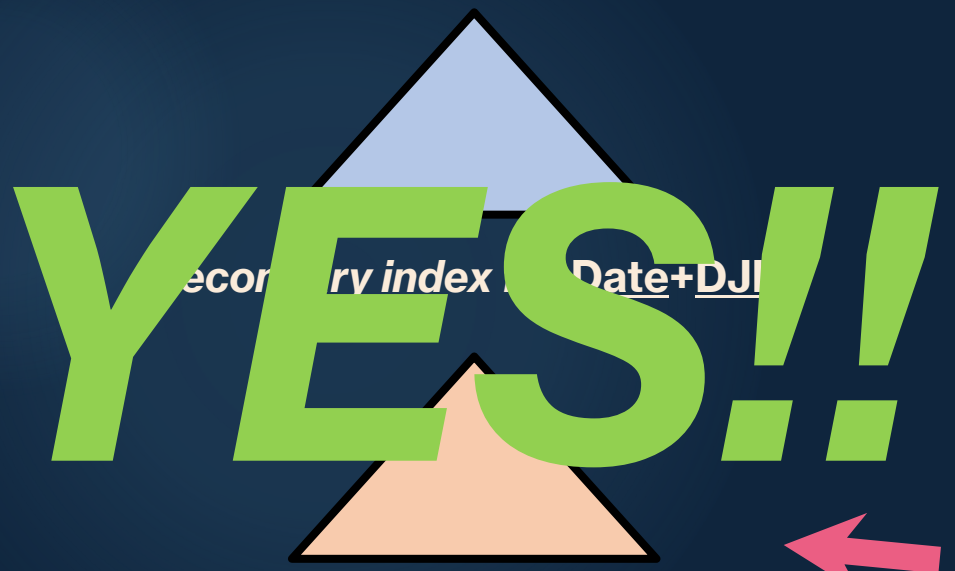
Can we avoid building another complete index?

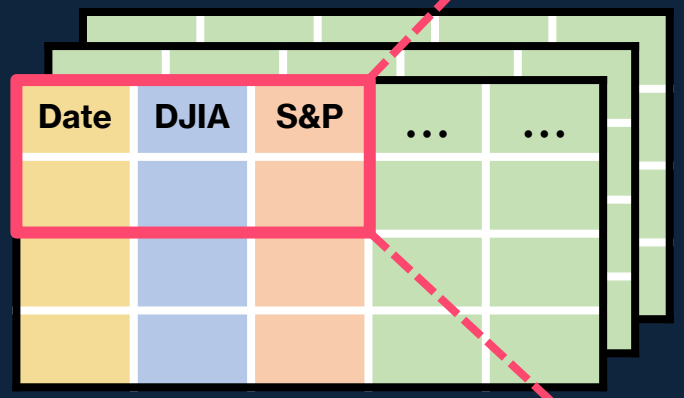
HERMIT: OBSERVATION

Column correlations **are prevalent**

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)





STOCK *table*

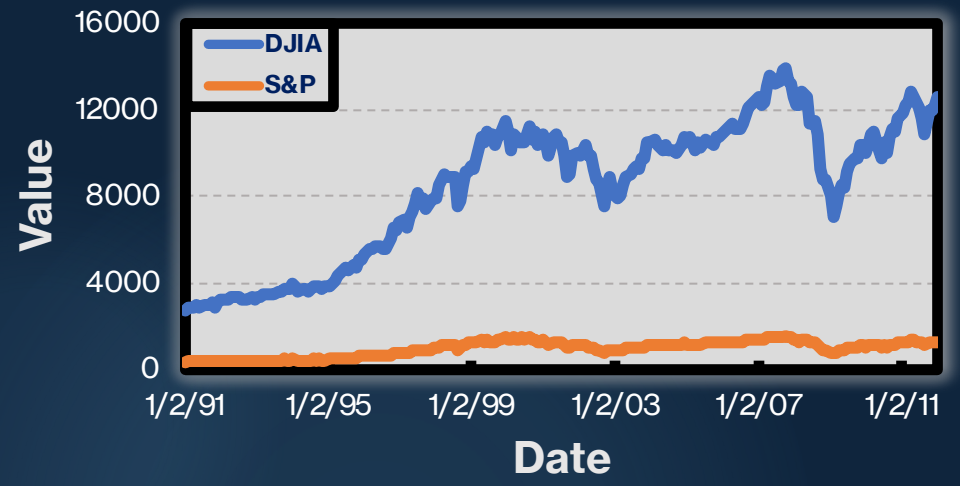
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

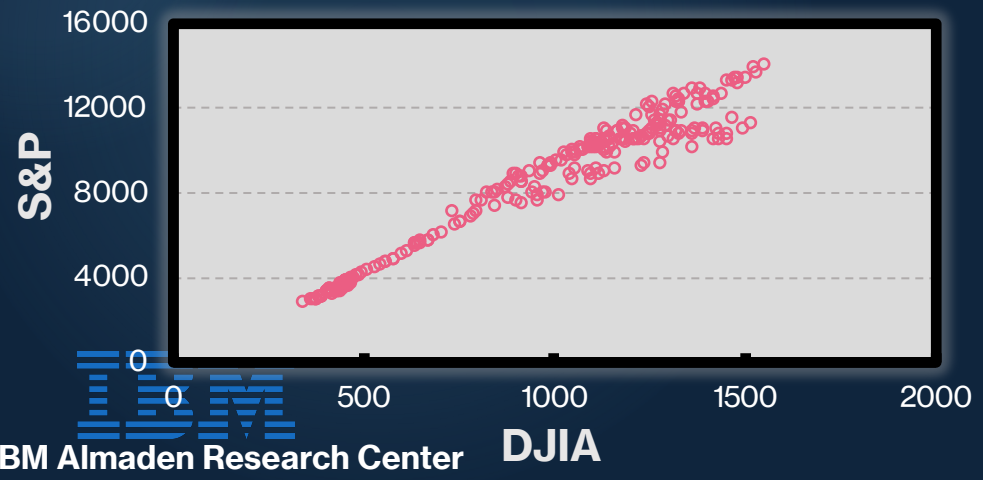
Secondary index for Date+S&P

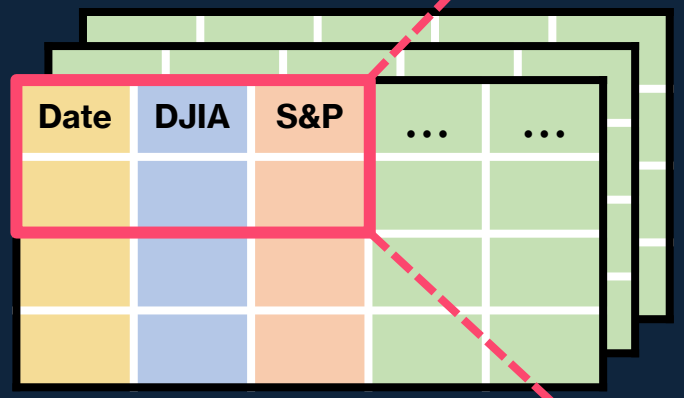
Can we avoid building another complete index?

HERMIT: OBSERVATION

DJIA and S&P are *highly correlated*





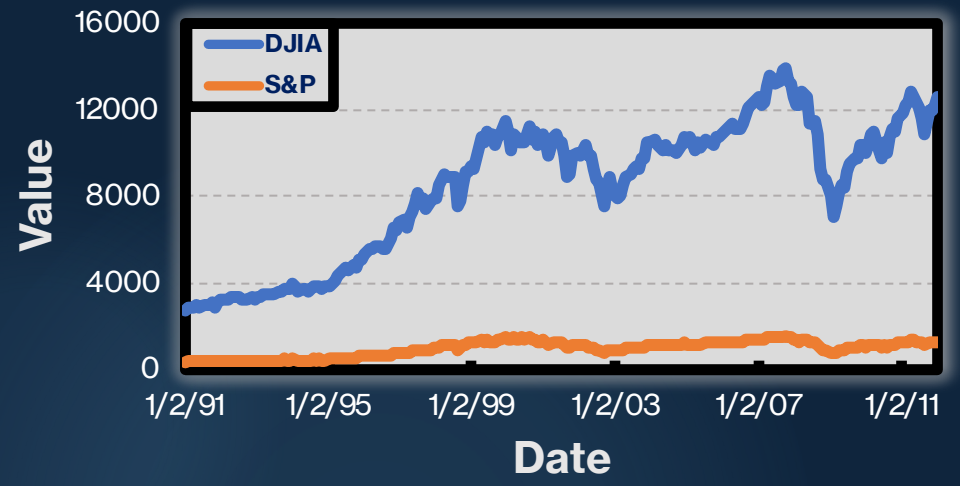


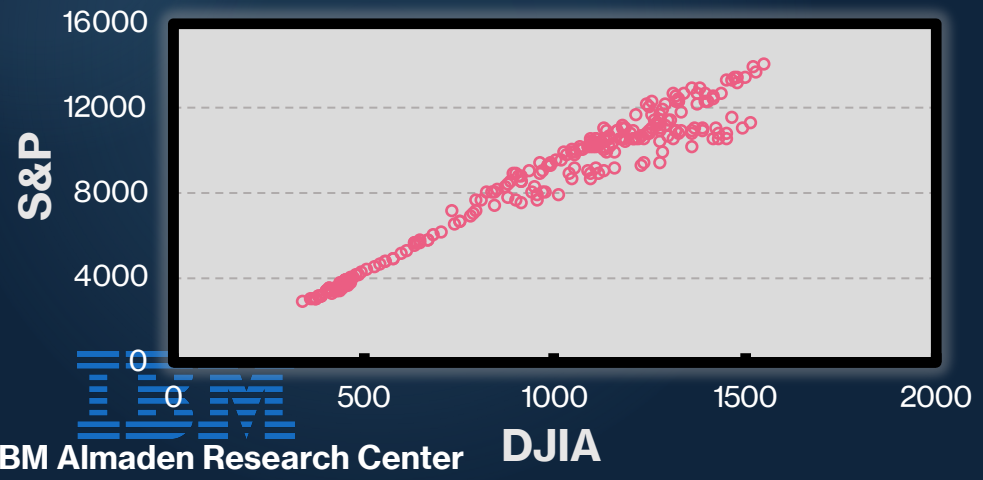
STOCK *table*

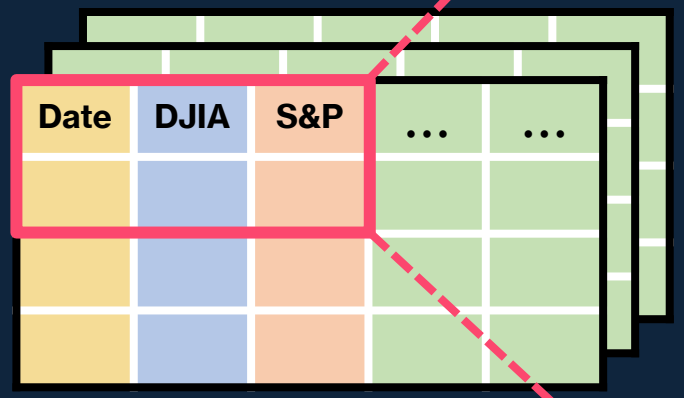
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: IDEA

Model the correlation between DJIA and S&P







STOCK *table*

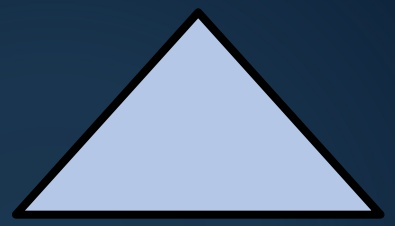
Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: IDEA

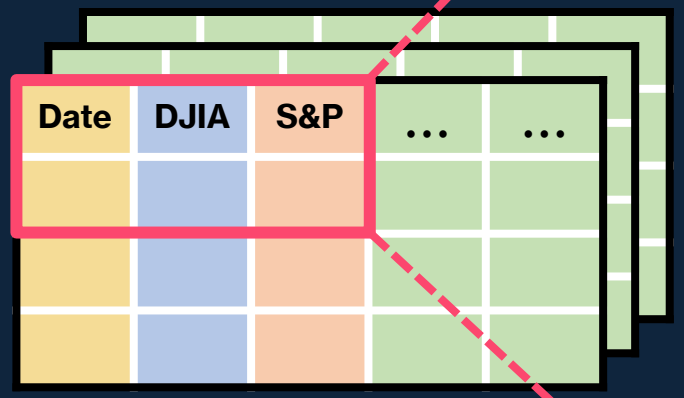
Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA



STOCK table

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

?

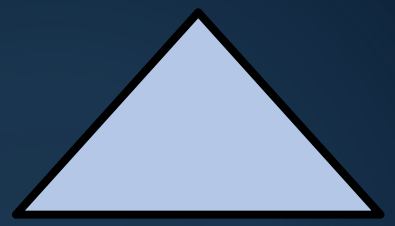
Secondary index for Date+S&P

HERMIT: IDEA

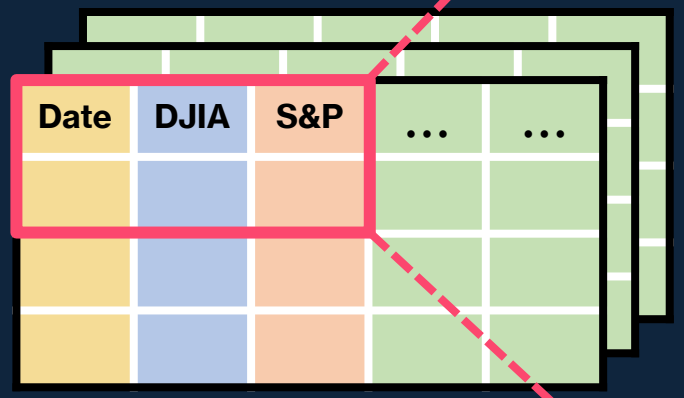
Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA



STOCK table

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

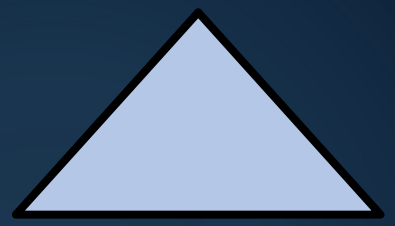
***succinct** index for S&P -> model correlations*

HERMIT: IDEA

Model the correlation between DJIA and S&P

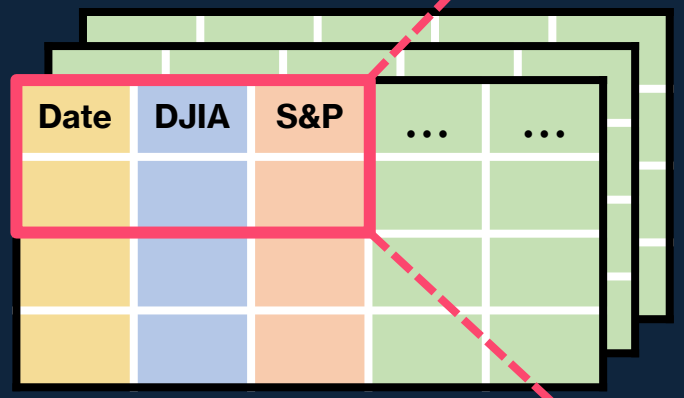
QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA

Input S&P value range



STOCK table

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

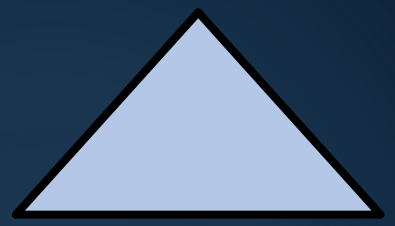
succinct index for S&P -> model correlations

HERMIT: IDEA

Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

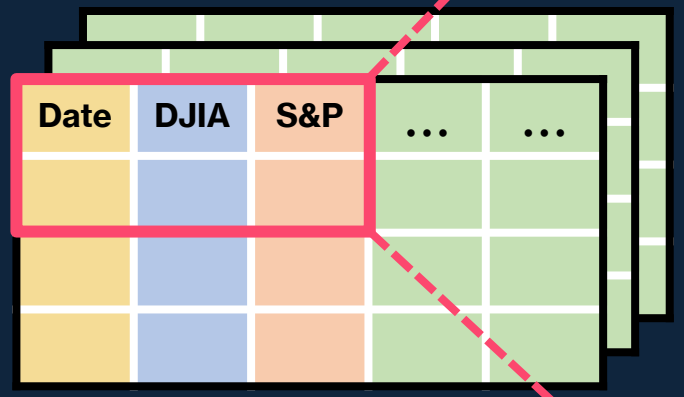
QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Secondary index for Date+DJIA

Input S&P value range

Output DJIA value range



STOCK table

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

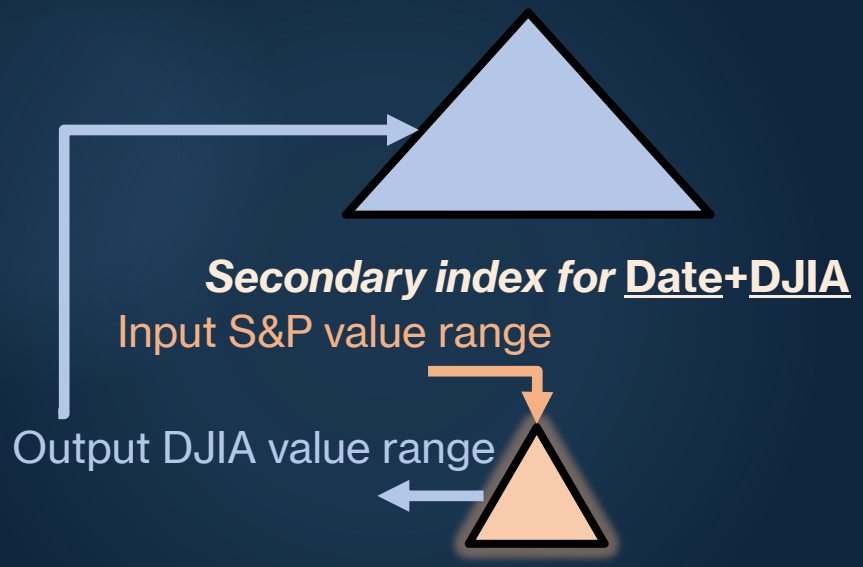
succinct index for S&P -> model correlations

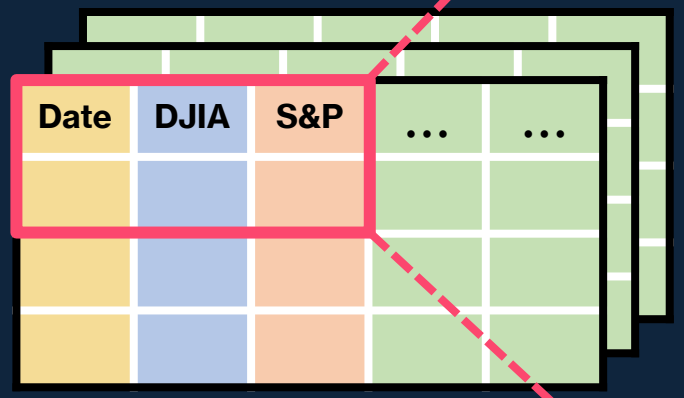
HERMIT: IDEA

Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)





STOCK *table*

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

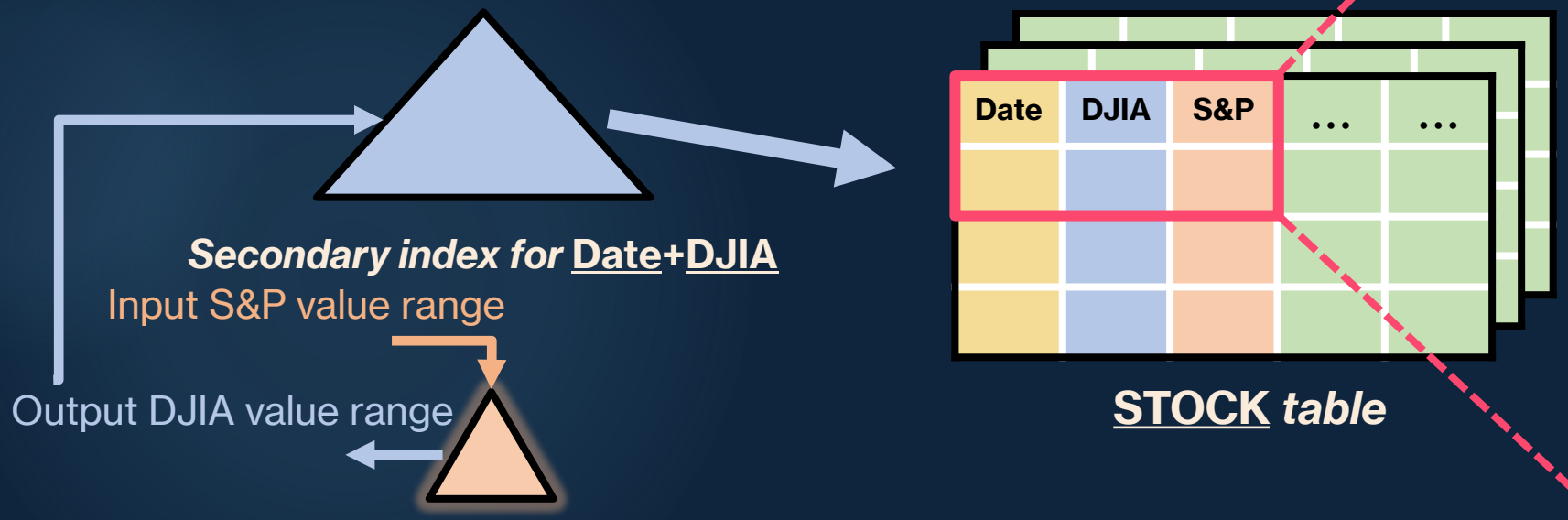
succinct *index for* S&P ***-> model correlations***

HERMIT: IDEA

Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (djia BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)



Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

***succinct** index for S&P -> model correlations*

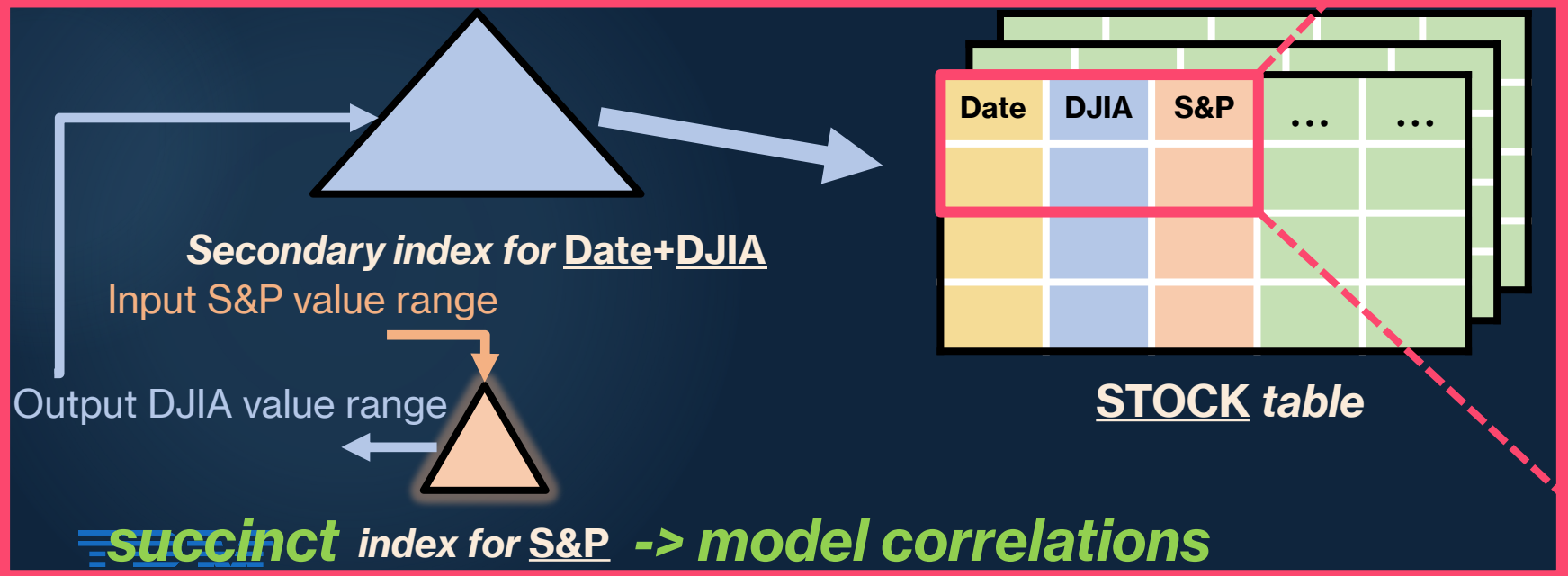
HERMIT: IDEA

Model the correlation between DJIA and S&P

QUERY 1: SELECT * FROM STOCK WHERE (date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)

QUERY 2: SELECT * FROM STOCK WHERE (date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)

HERMIT



Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
5/1/92	3396.90	415.35
6/1/92	3318.50	408.14
7/1/92	3393.80	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71

HERMIT: IDEA

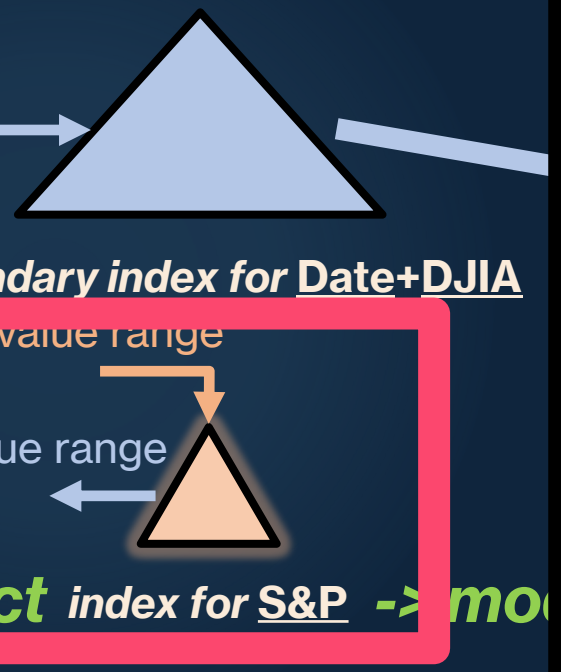
Model the correlation between DJIA and S&P

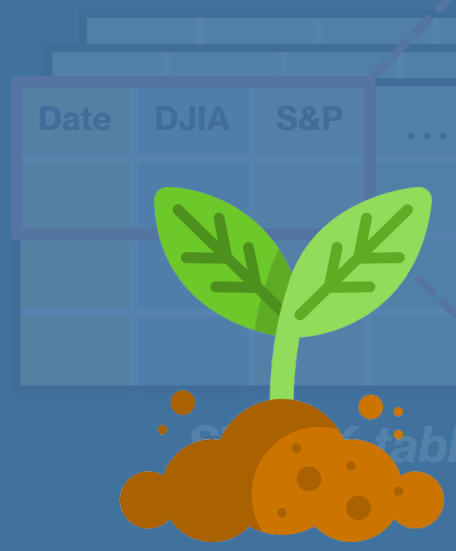
QUERY 1: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)

QUERY 2: SELECT * FROM stock WHERE
(date BETWEEN ? AND ?) AND (S&P BETWEEN ? AND ?)

HERMIT

Date	DJIA	S&P
1/2/91	2736.39	343.93
2/1/91	2882.18	367.07
3/1/91	2913.86	375.22
4/1/91	2887.87	375.34
5/1/91	3027.50	389.83
6/3/91	2906.75	371.16
7/1/91	3024.82	387.81
8/1/91	3043.60	395.43
9/3/91	3016.77	387.86
10/1/91	3069.10	392.45
11/1/91	2894.68	375.22
12/2/91	3168.83	417.09
1/2/92	3223.40	408.78
2/3/92	3267.70	412.70
3/2/92	3235.50	403.69
4/1/92	3359.10	414.95
[illegible]	[illegible]	415.35
[illegible]	[illegible]	408.14
[illegible]	[illegible]	424.21
8/3/92	3257.40	414.03
9/1/92	3271.70	417.80
10/1/92	3226.30	418.68
11/2/92	3305.20	431.35
12/1/92	3301.11	435.71





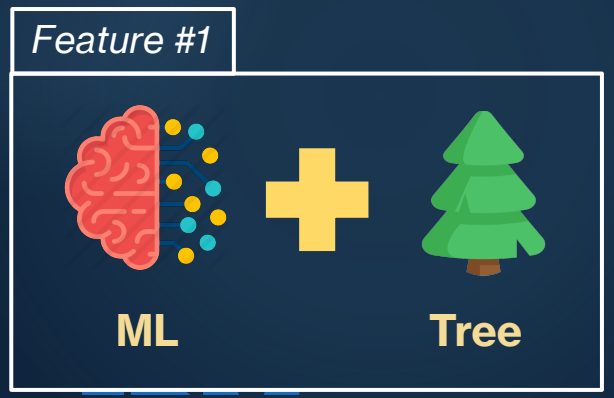
TRS-TREE

HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)

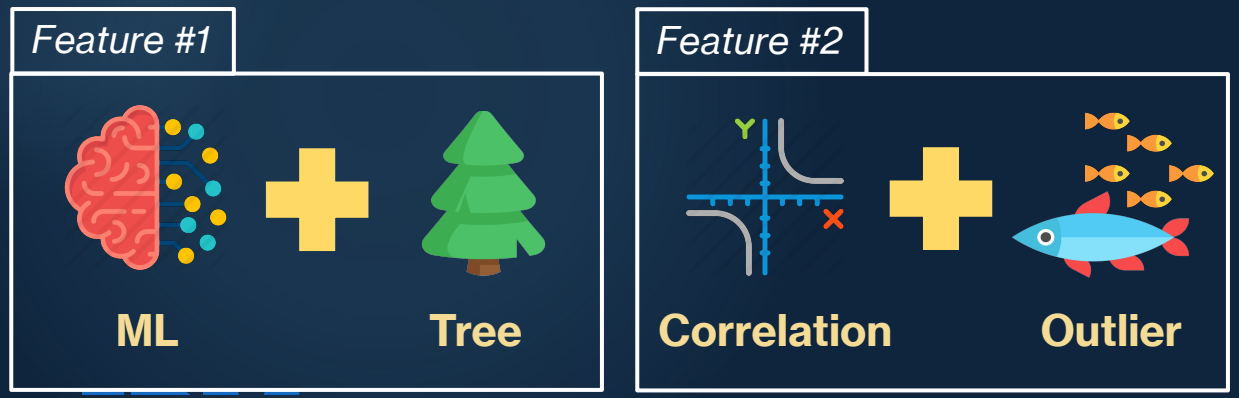
HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)
  - A succinct, ***ML-enhanced***, tree-structured index



HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)
  - A succinct, ***ML-enhanced***, tree-structured index
  - Capture ***correlation*** while handling ***outliers***



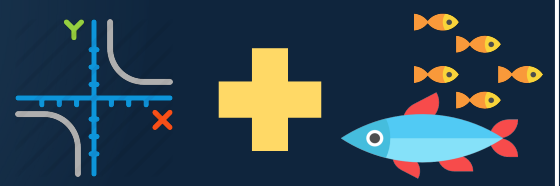
HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)
  - A succinct, ***ML-enhanced***, tree-structured index
  - Capture ***correlation*** while handling ***outliers***
  - ***Adaptively*** and ***dynamically*** construct and reorganize internal structures

Feature #1

ML **Tree**

Feature #2



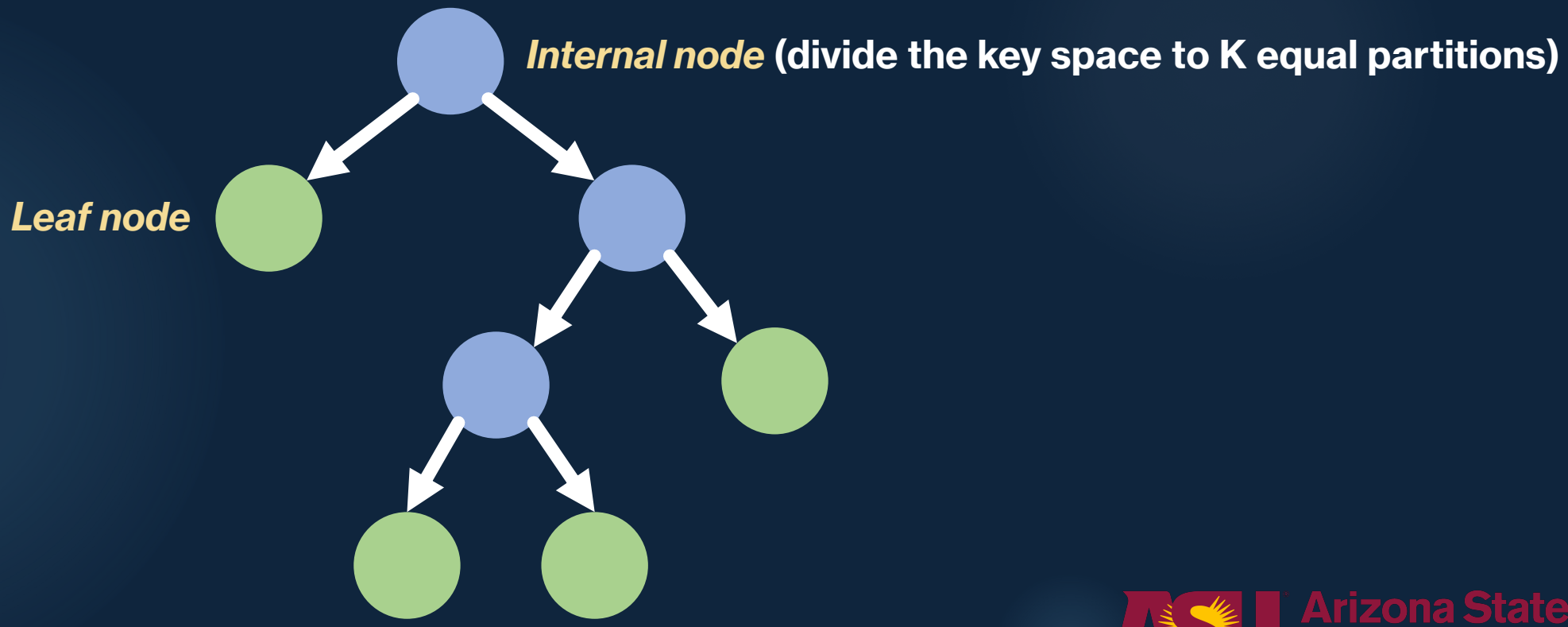
Correlation **Outlier**

Feature #3

Adaptive **Dynamic**

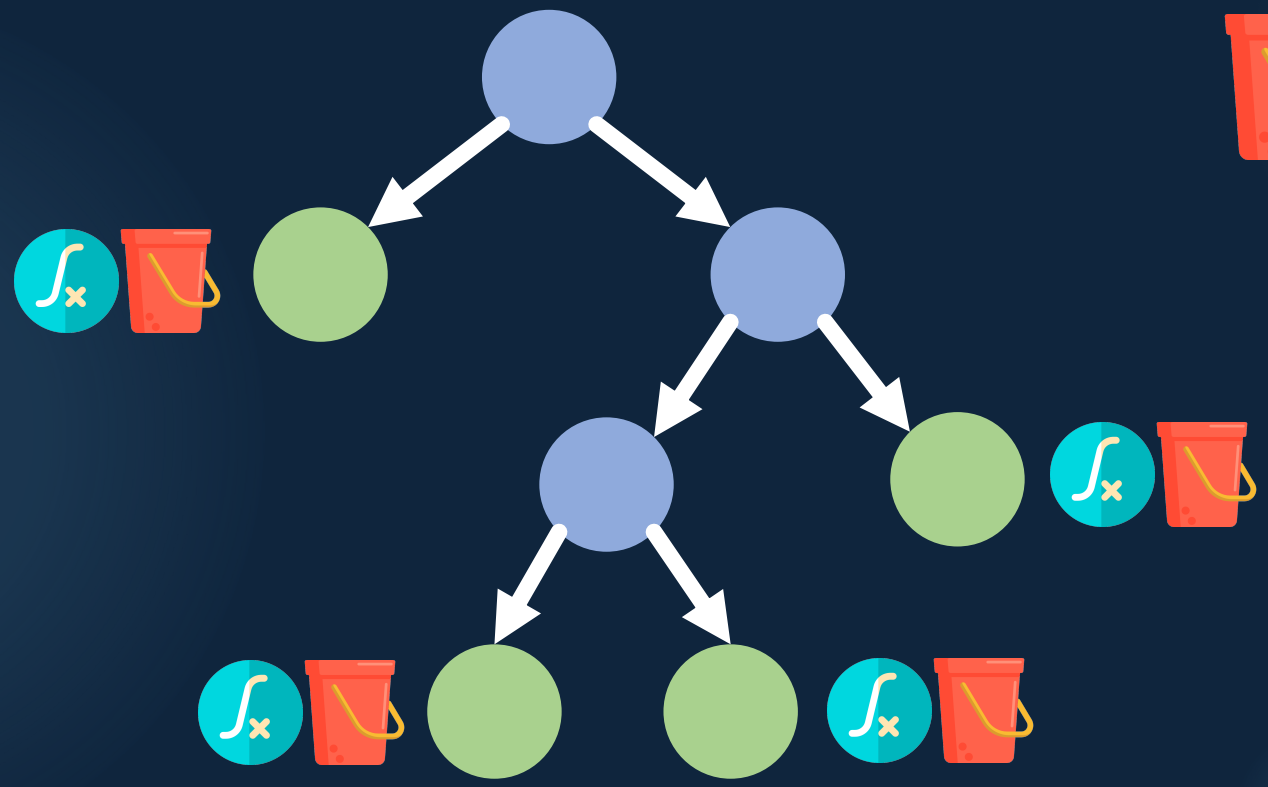
HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)



HERMIT: TRS-TREE DESIGN

- Tiered Regression Search Tree (TRS-Tree)

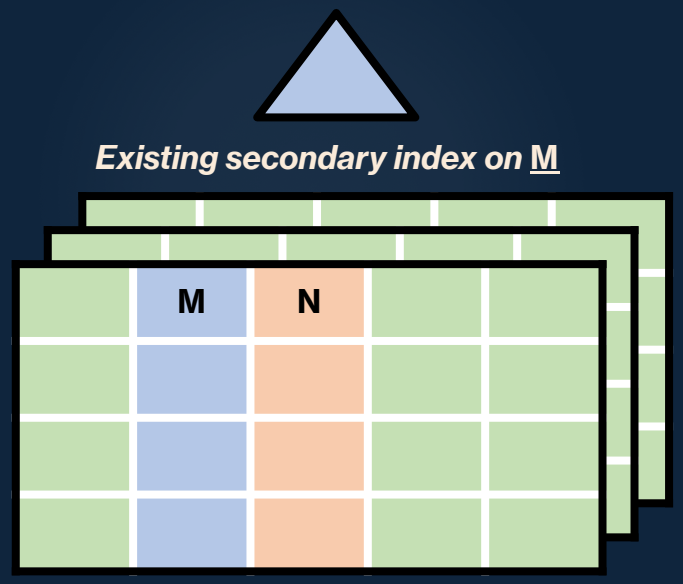


Correlation function
$y = \beta x + \alpha \pm \epsilon$
(simple linear regression)

Outlier buffer

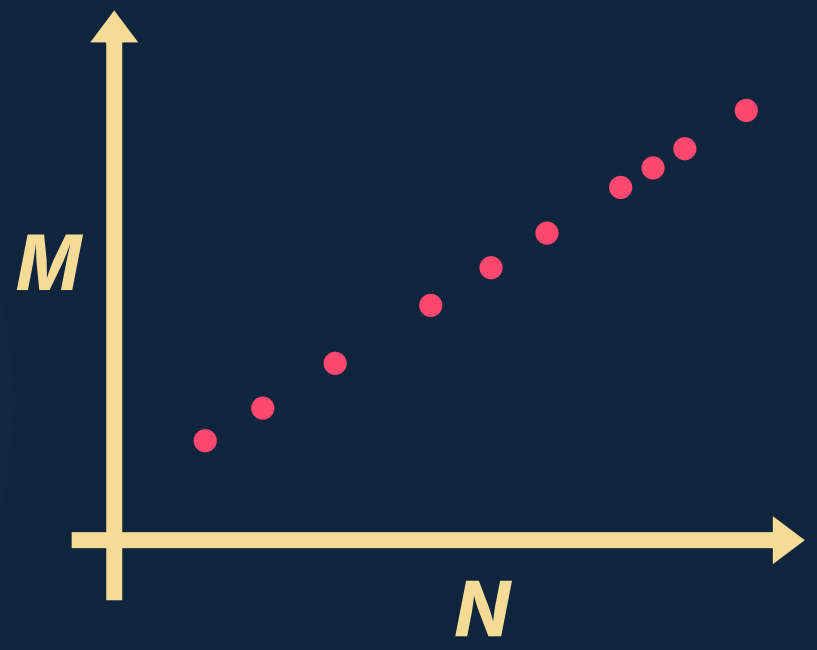
HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

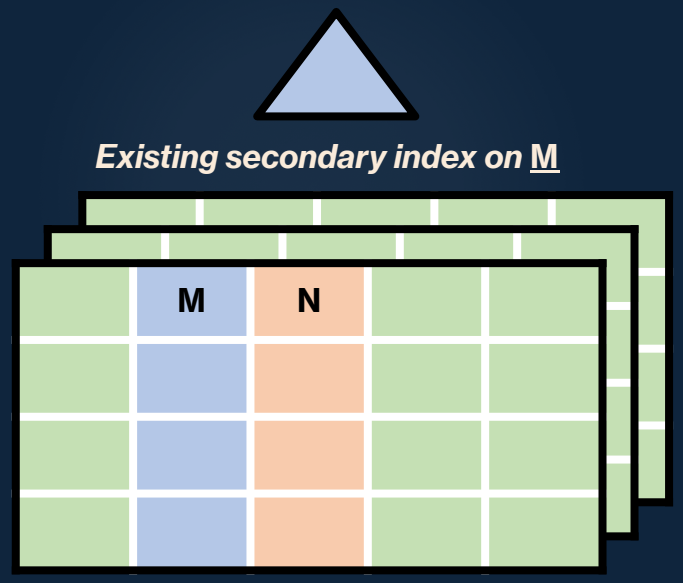


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

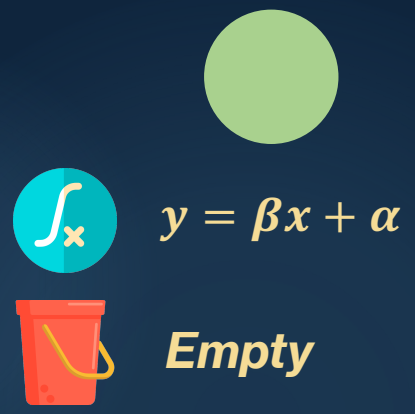


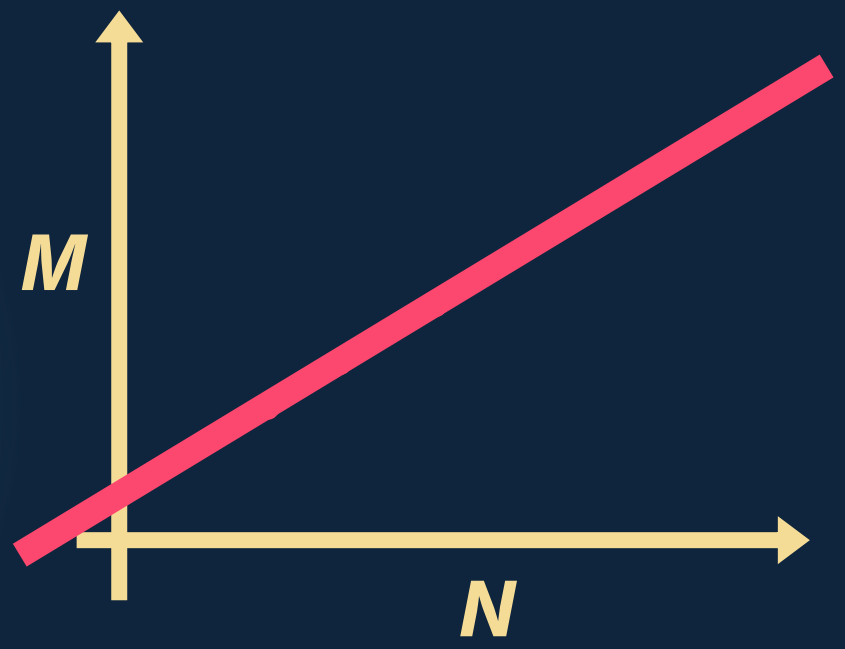
TRS-Tree as a mapping from N to M

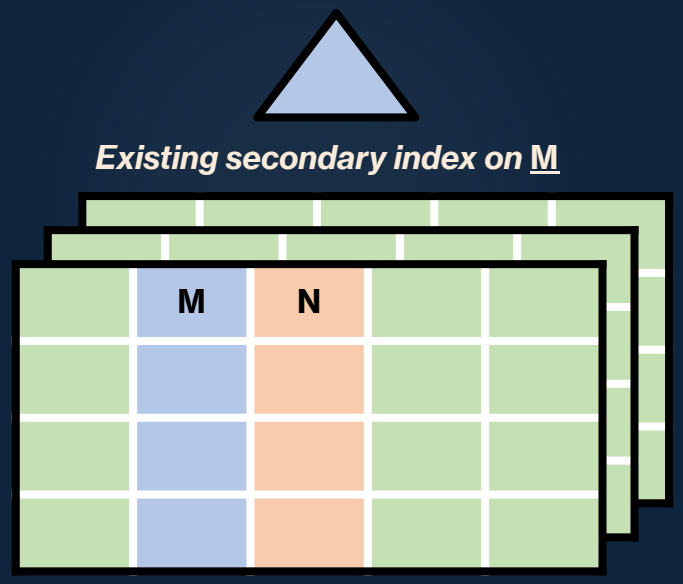


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

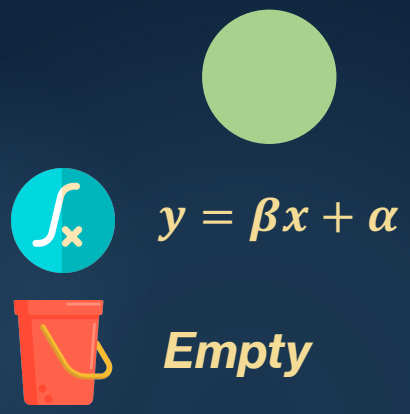


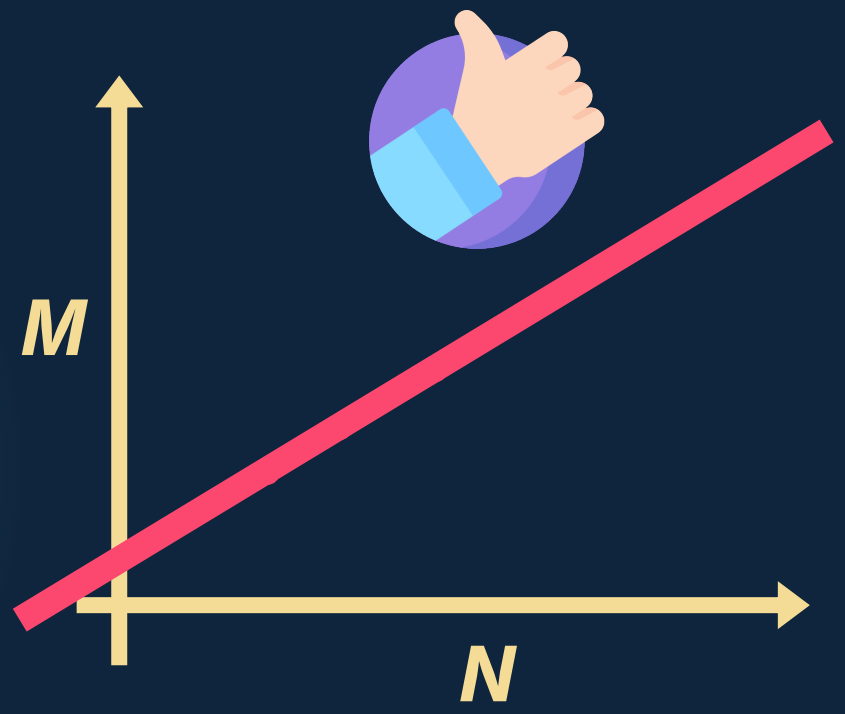


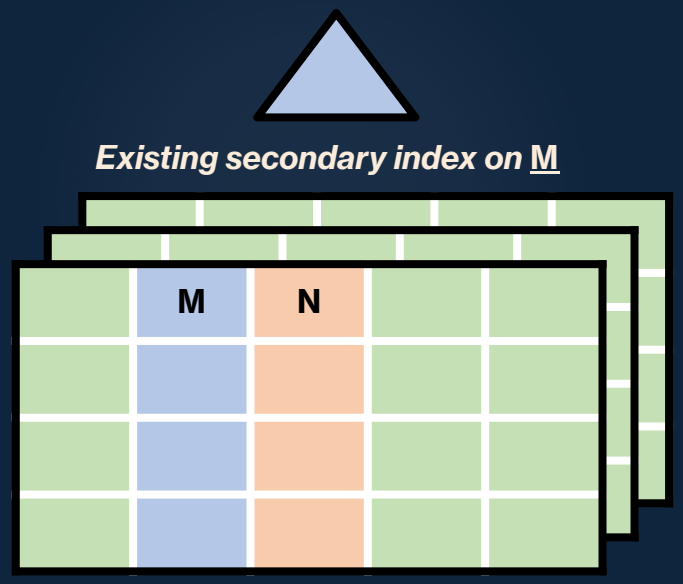


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

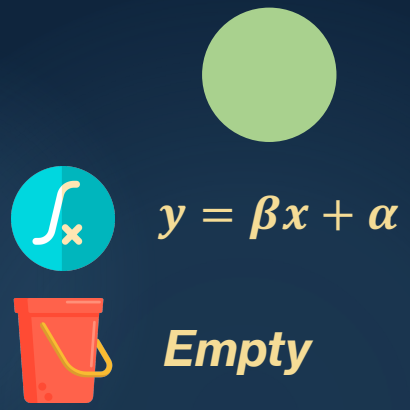


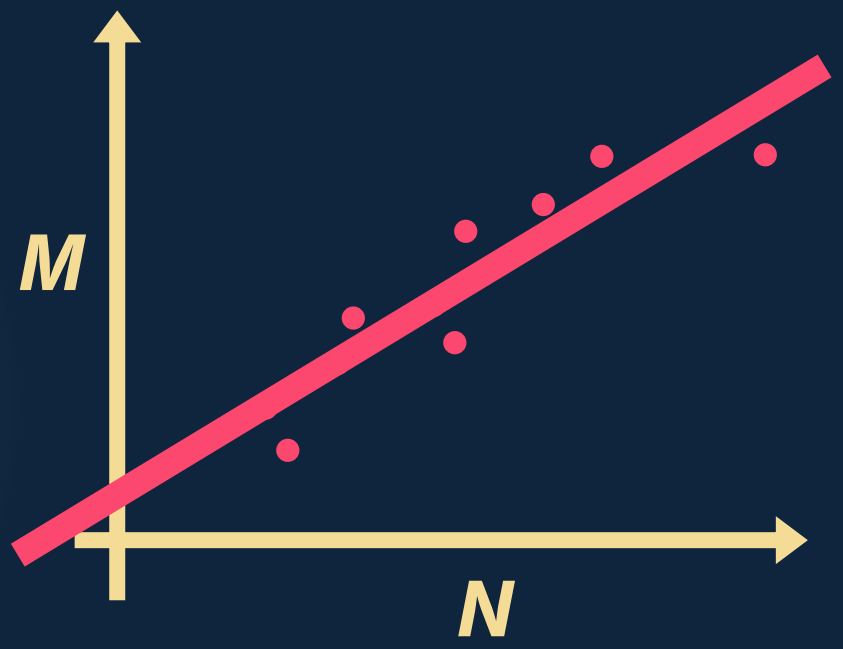


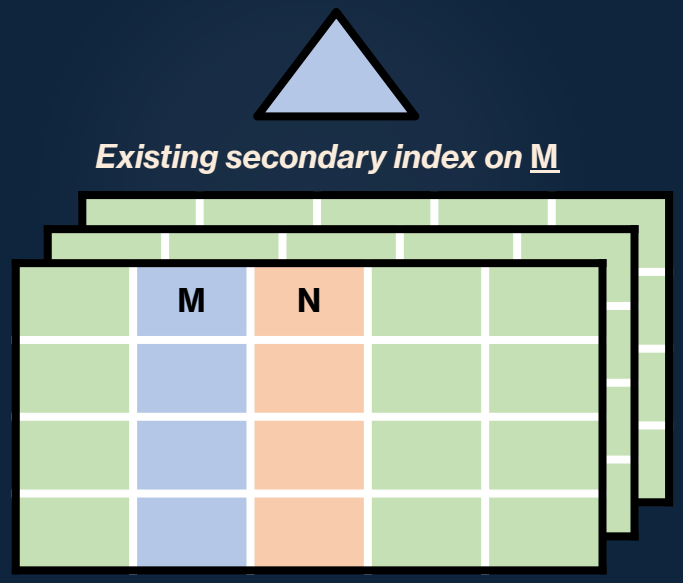


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

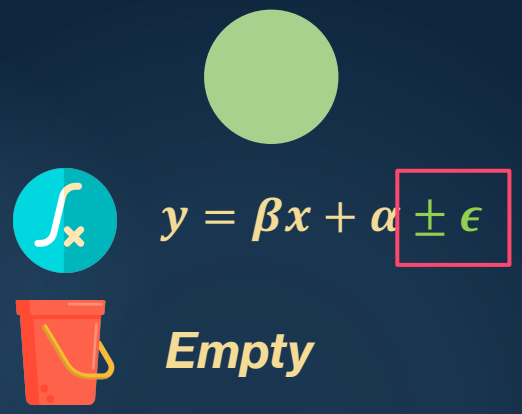


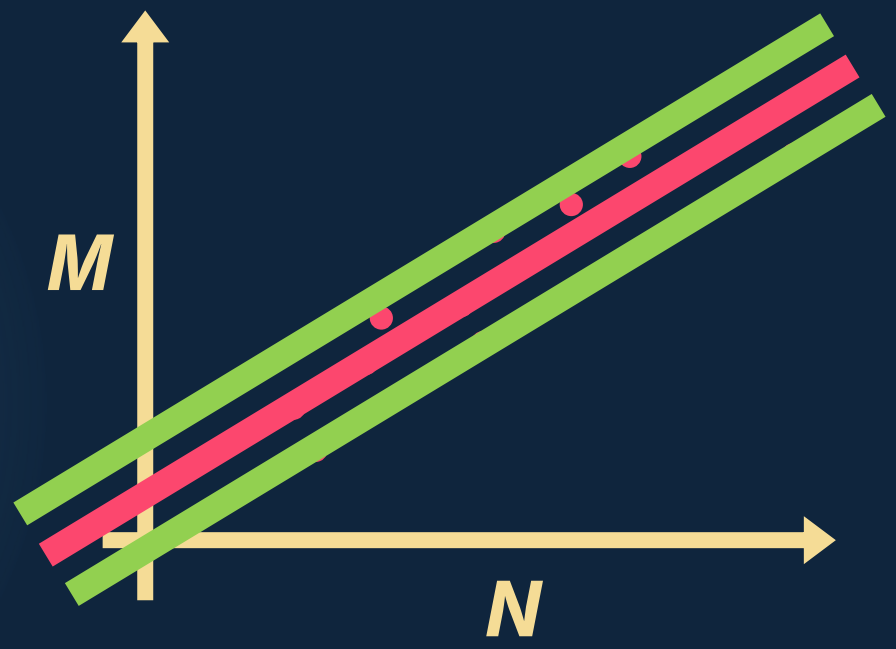


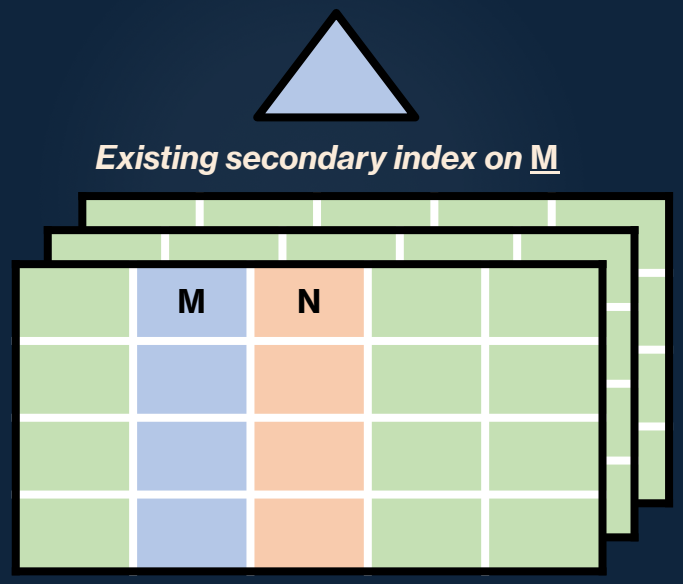


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

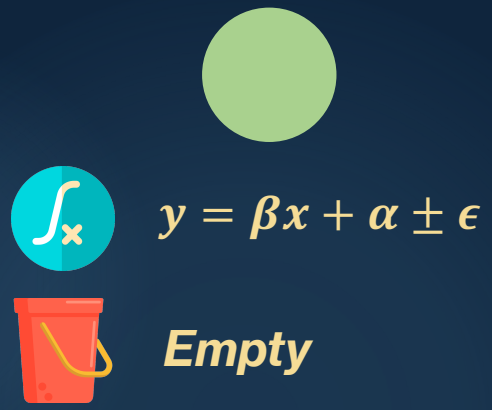


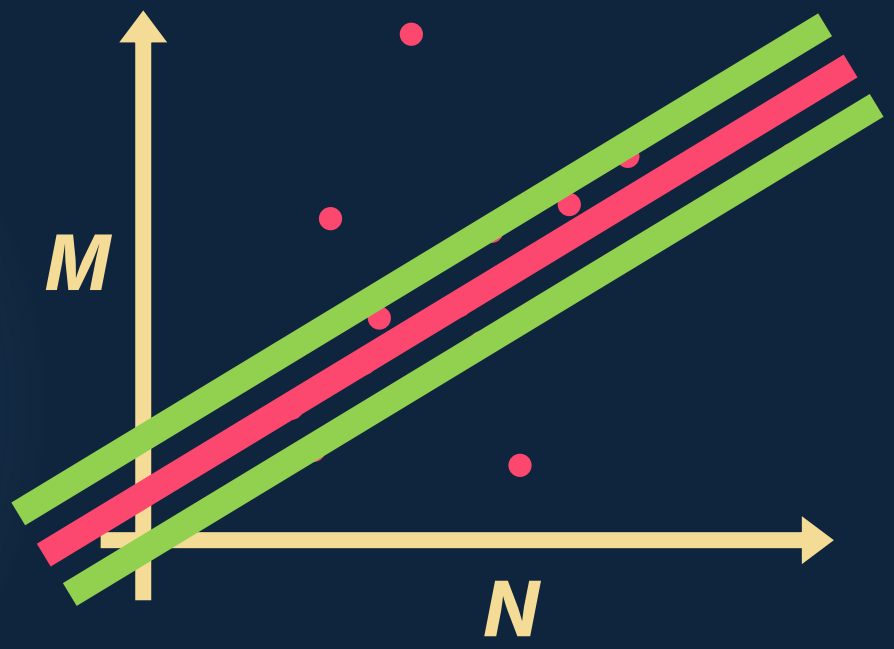


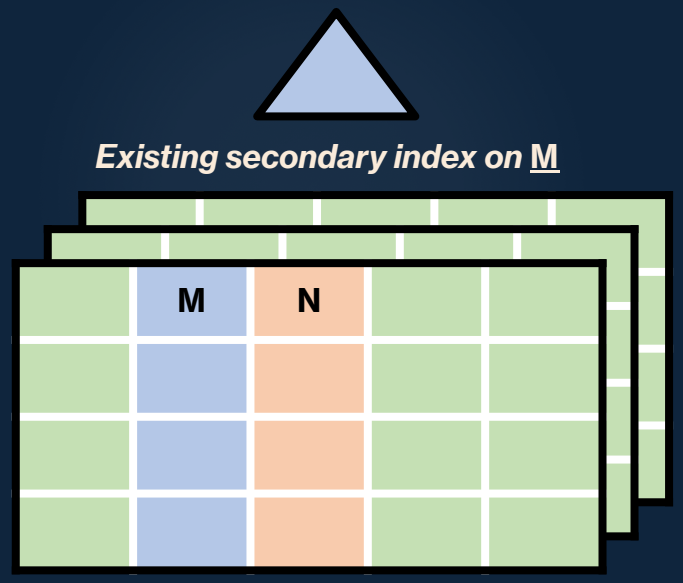


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

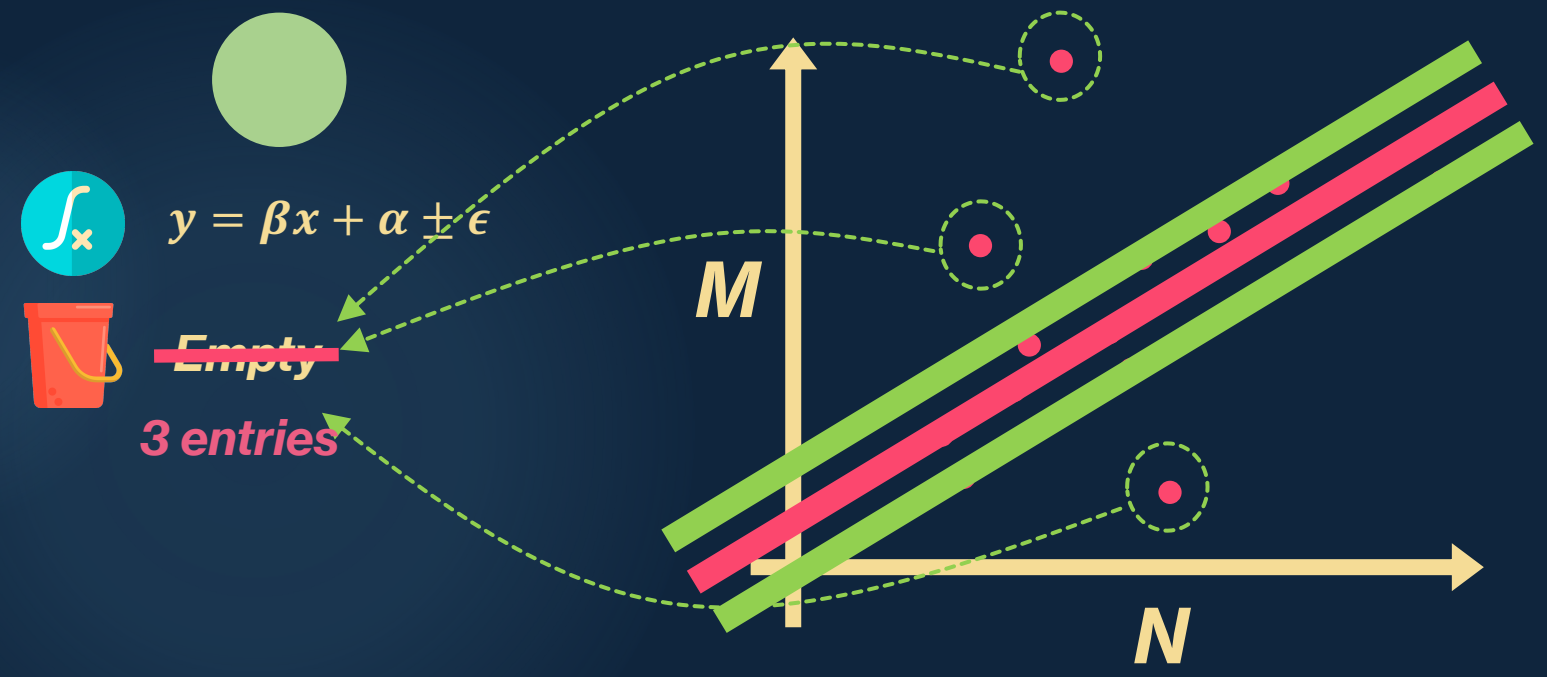


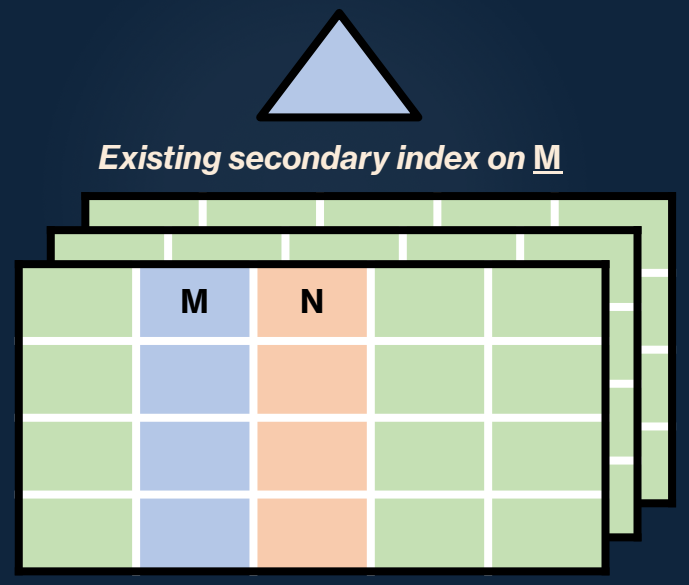




HERMIT: TRS-TREE DESIGN

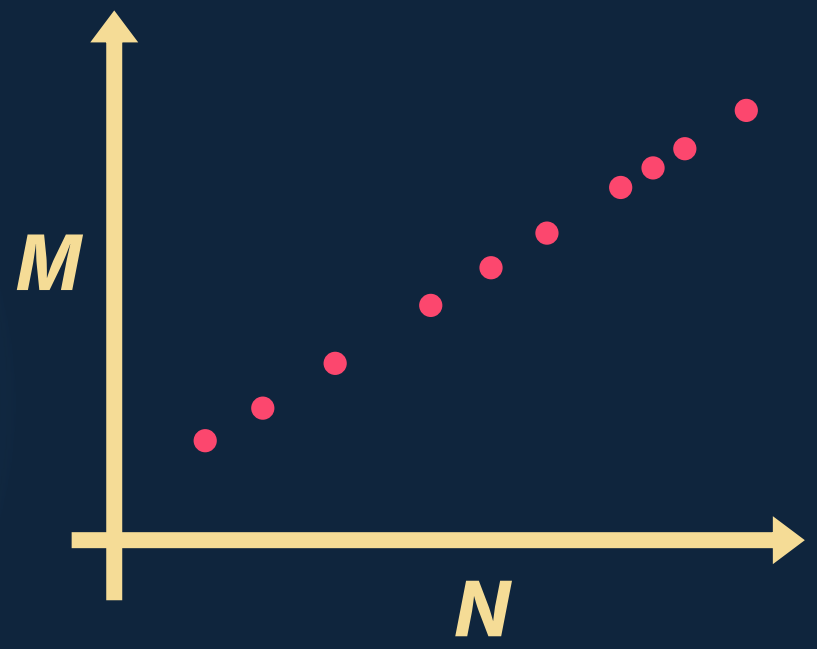
- TRS-Tree: model correlation between columns **M** and **N**

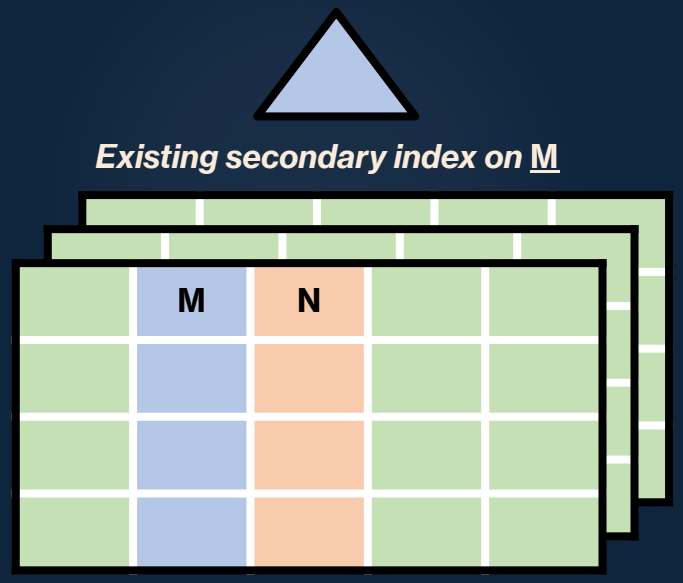




HERMIT: TRS-TREE DESIGN

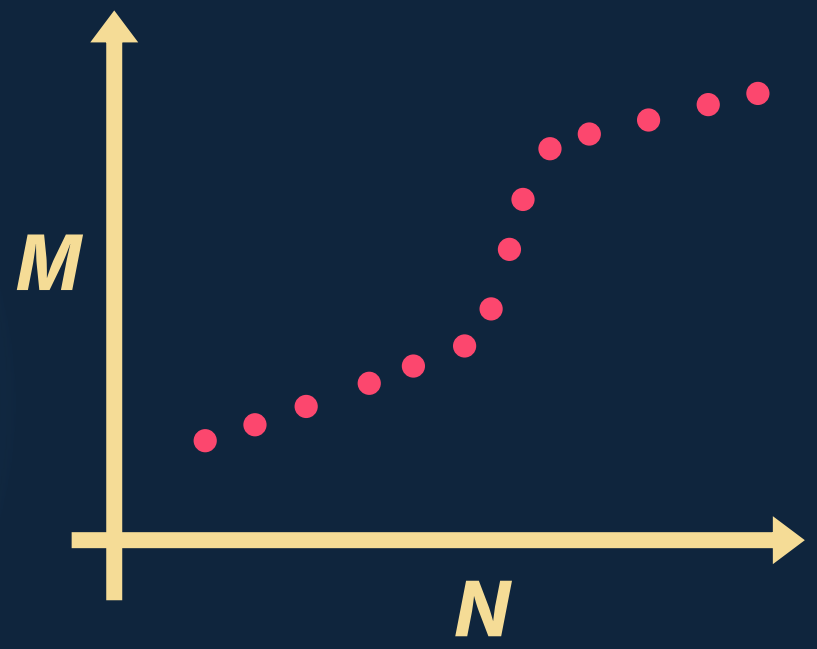
- TRS-Tree: model correlation between columns **M** and **N**

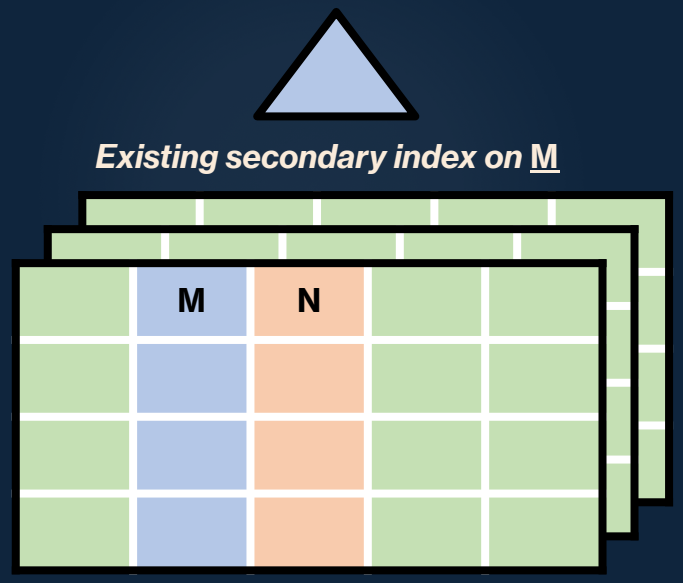




HERMIT: TRS-TREE DESIGN

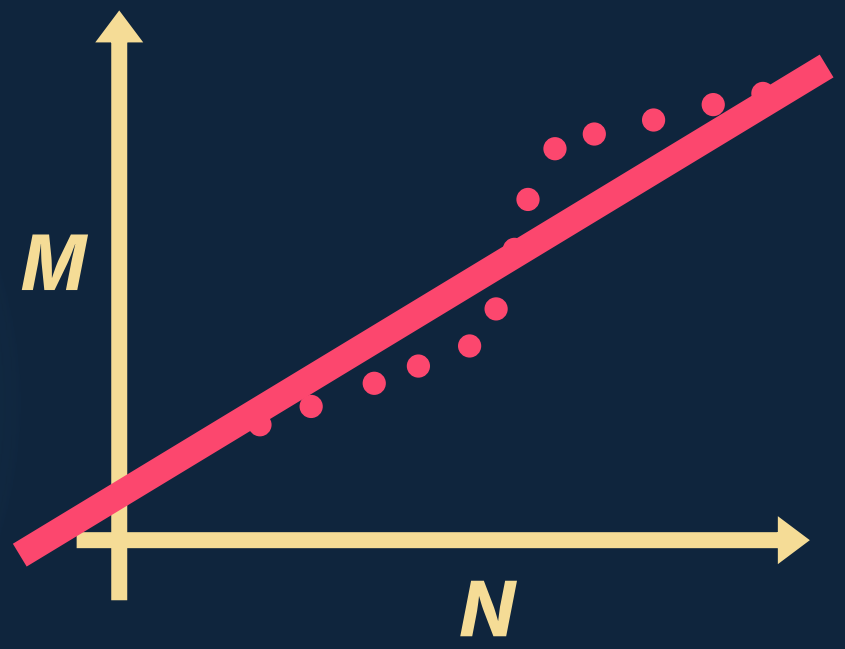
- TRS-Tree: model correlation between columns **M** and **N**

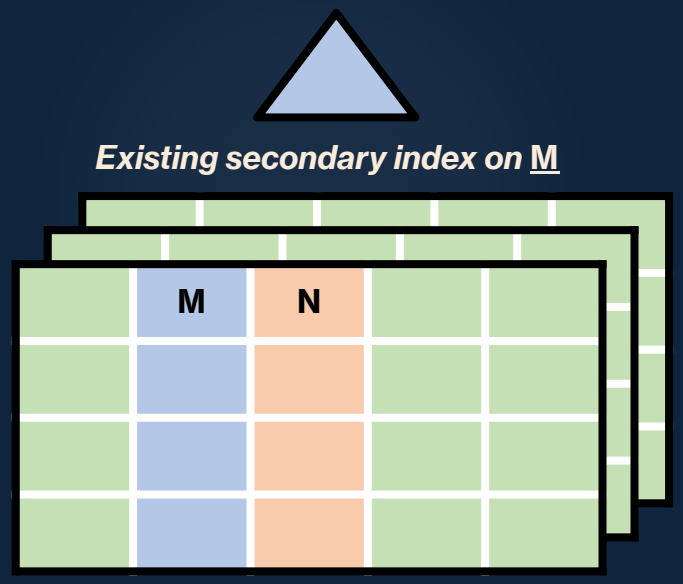




HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**



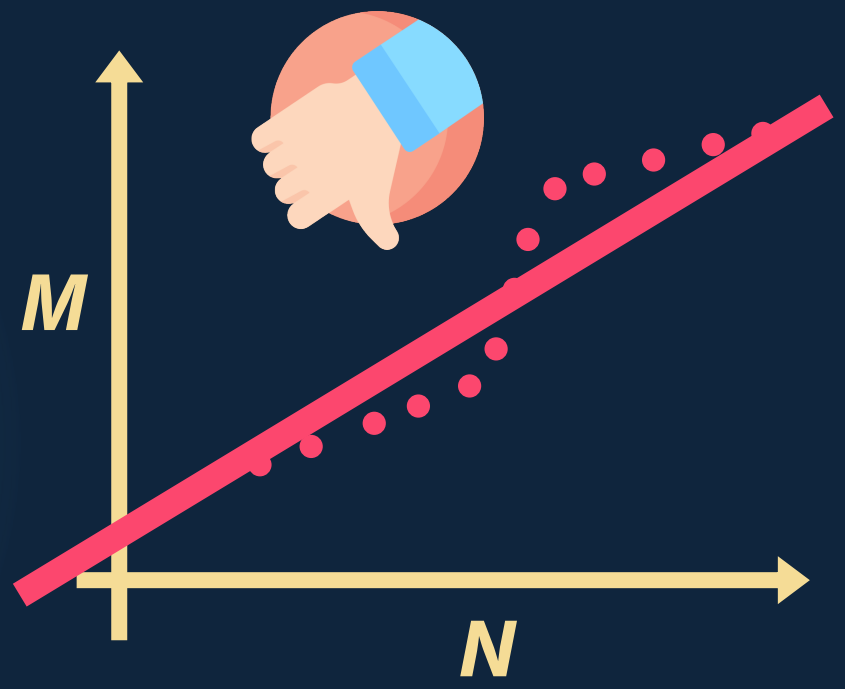


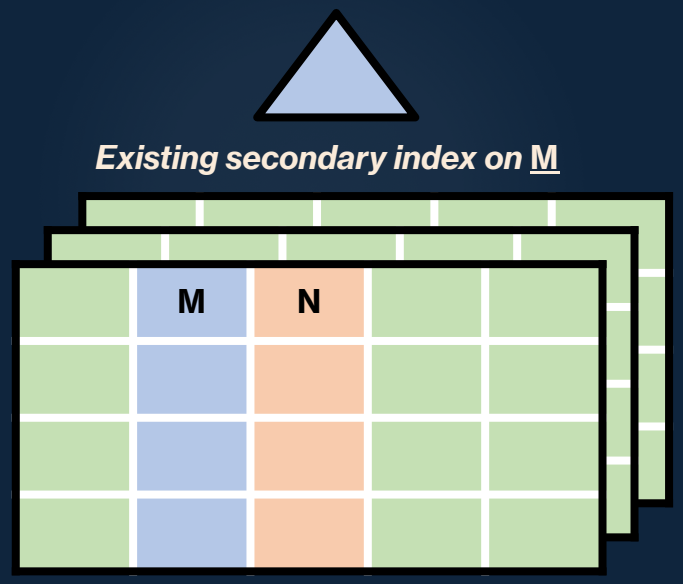
HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

$$y = \beta x + \alpha \pm \epsilon$$

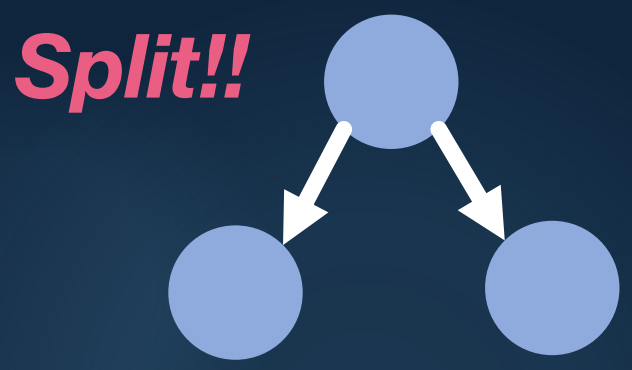
Lots of entries!!

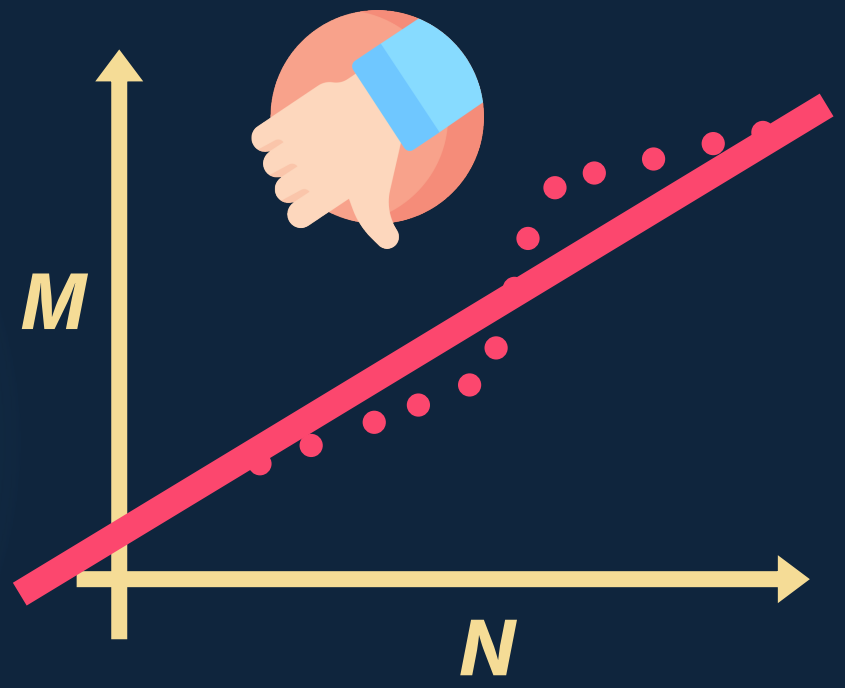


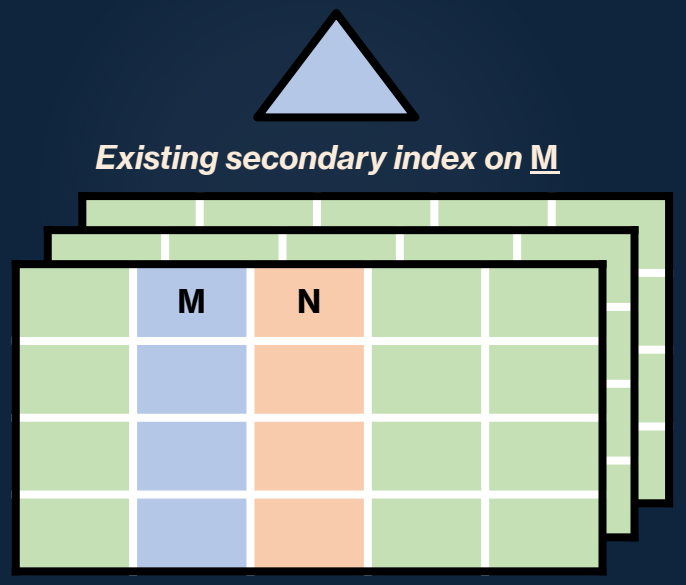


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

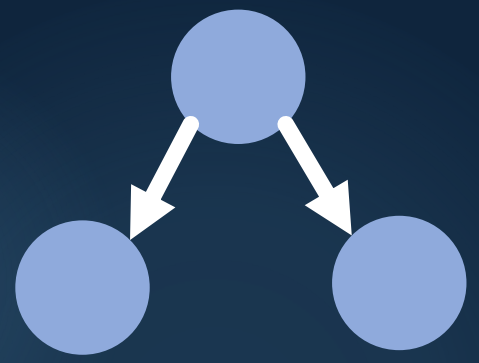


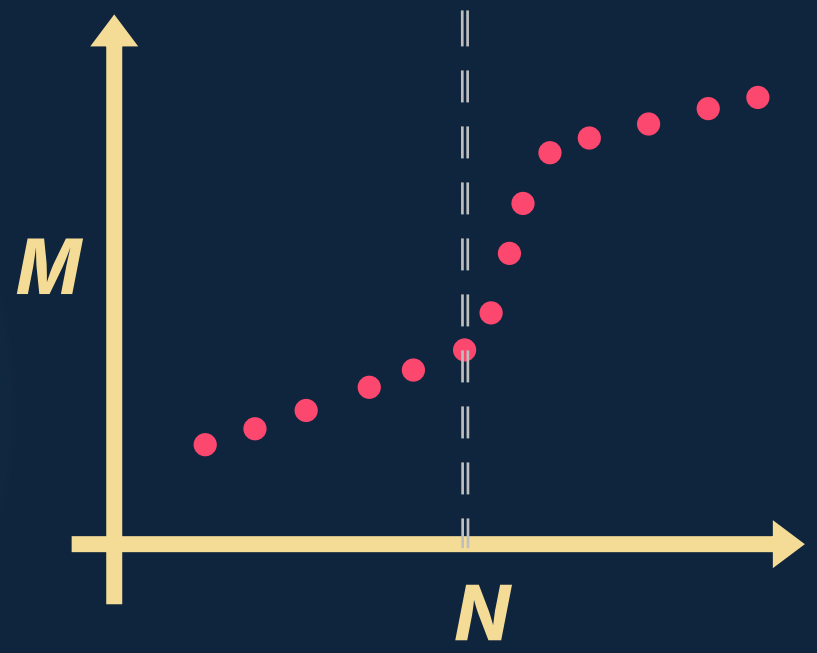


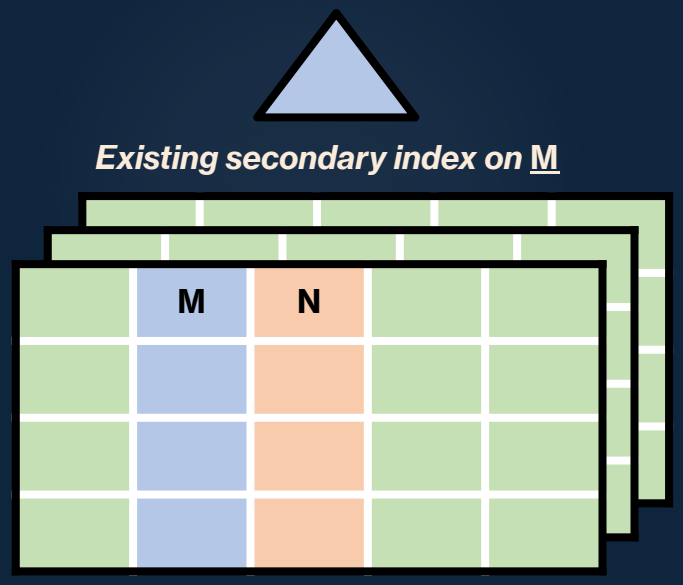


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

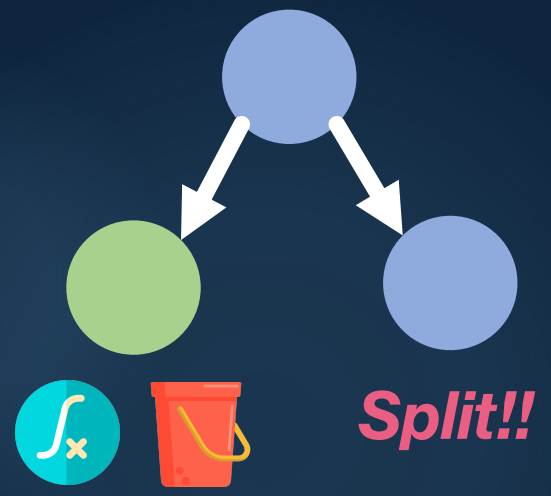


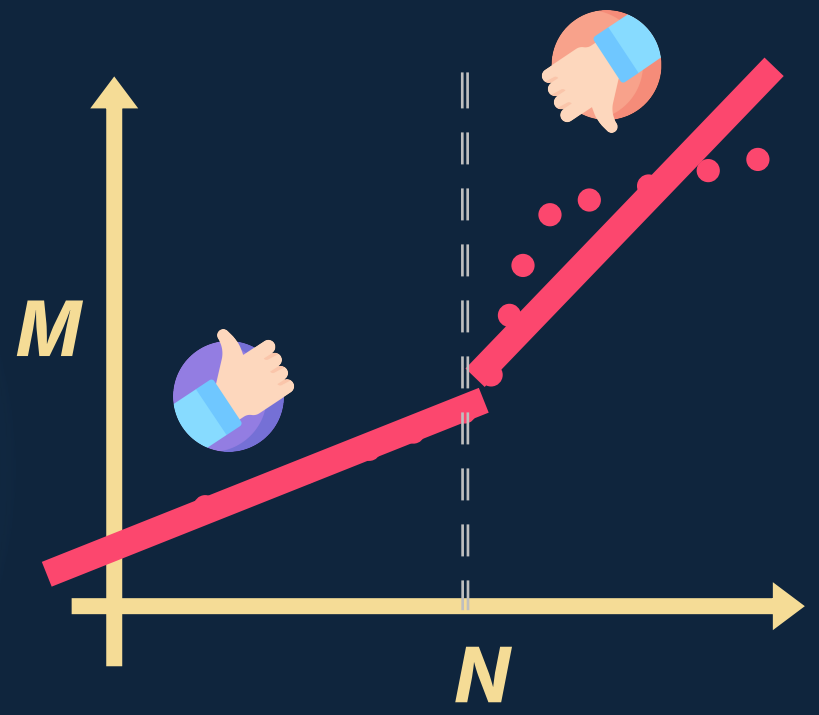


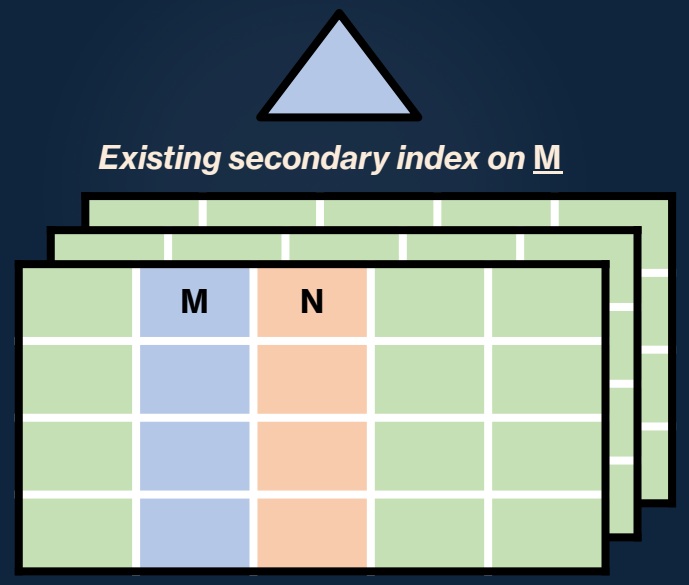


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

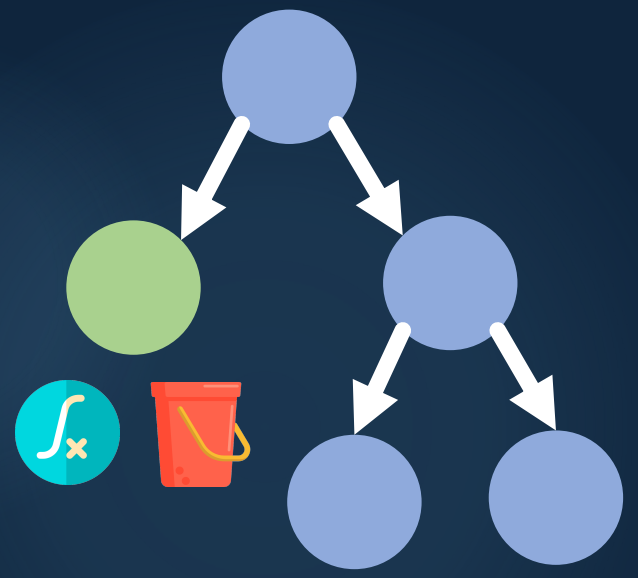


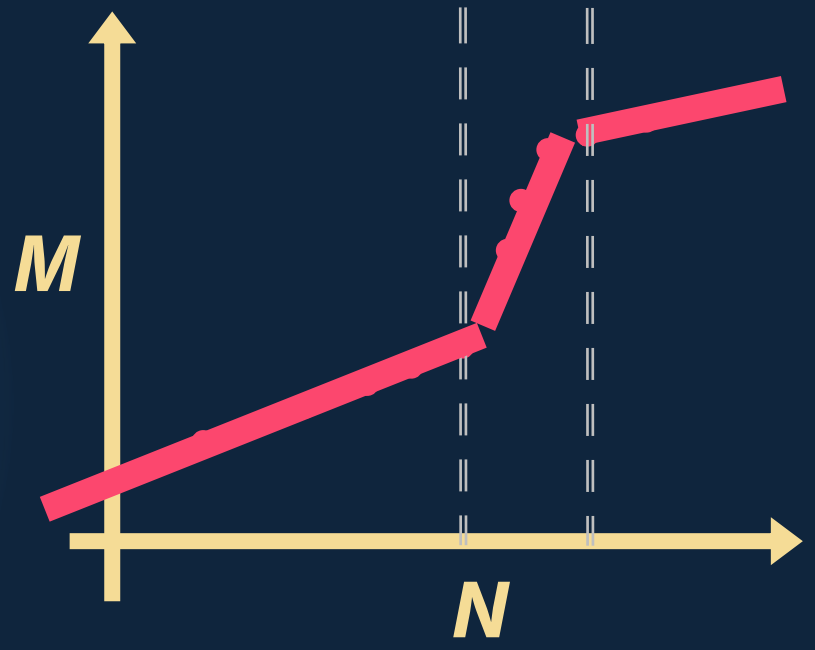


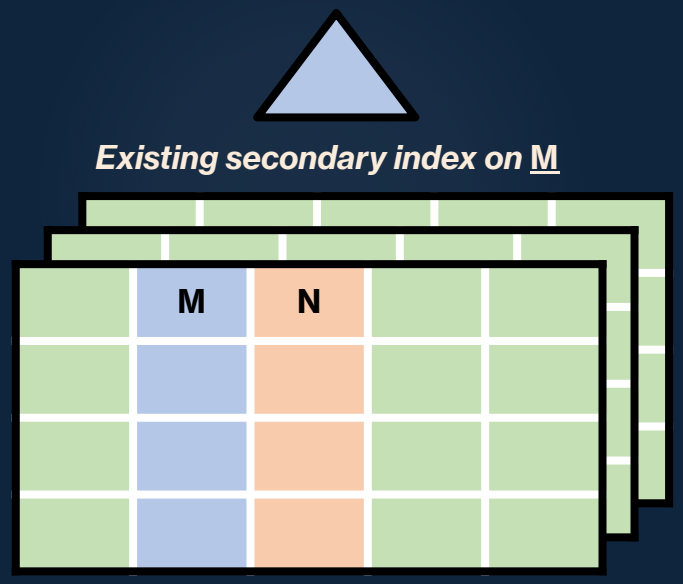


HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

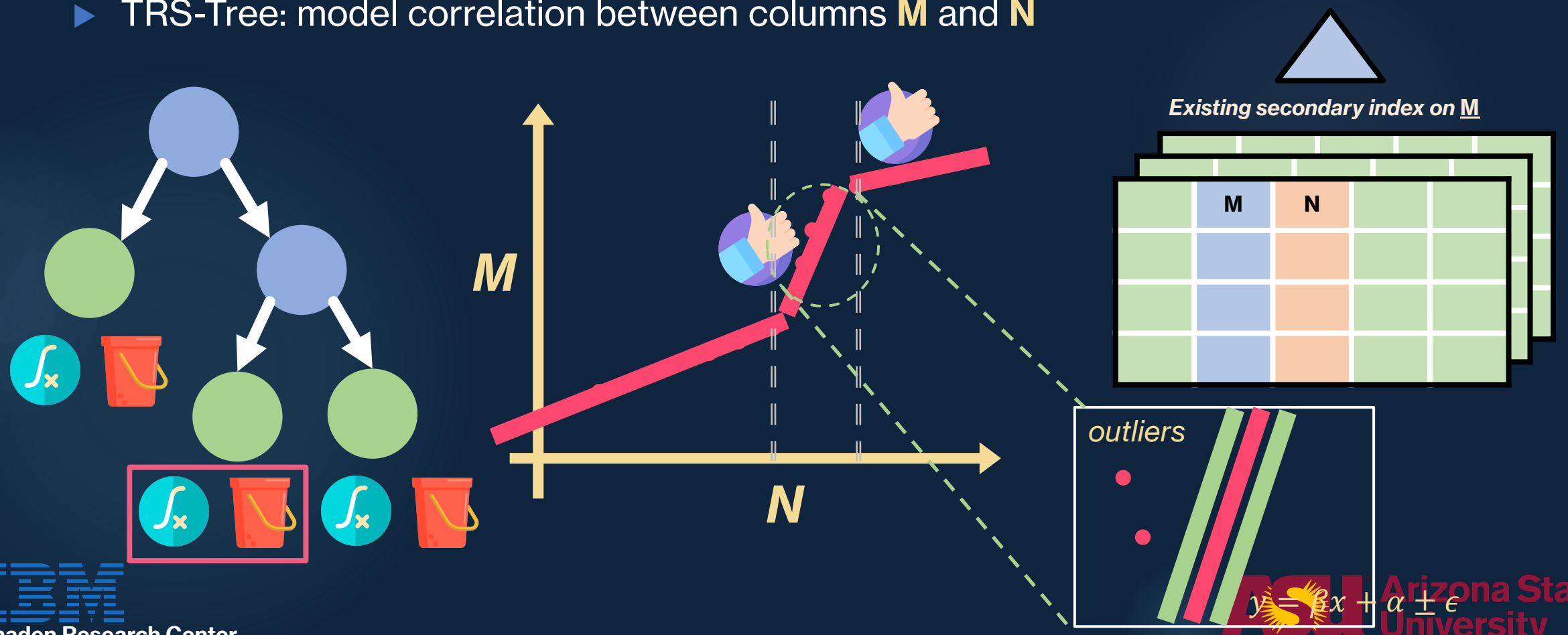






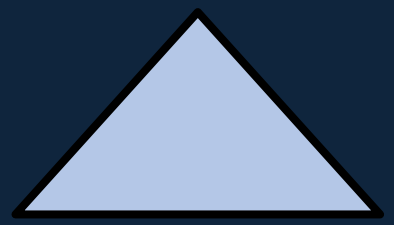
HERMIT: TRS-TREE DESIGN

- TRS-Tree: model correlation between columns **M** and **N**

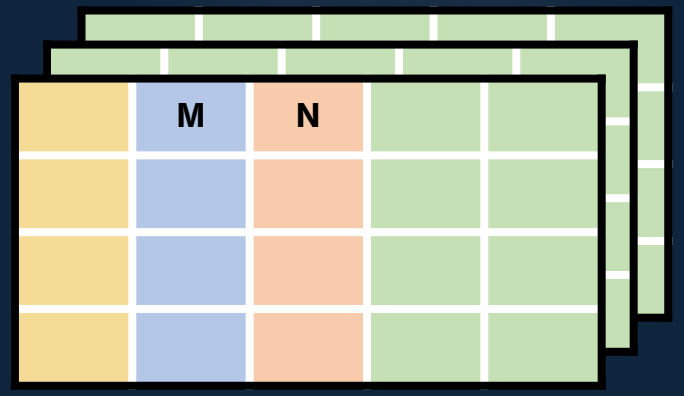


HERMIT: ACCESS METHOD

TRS-Tree Index on **N**

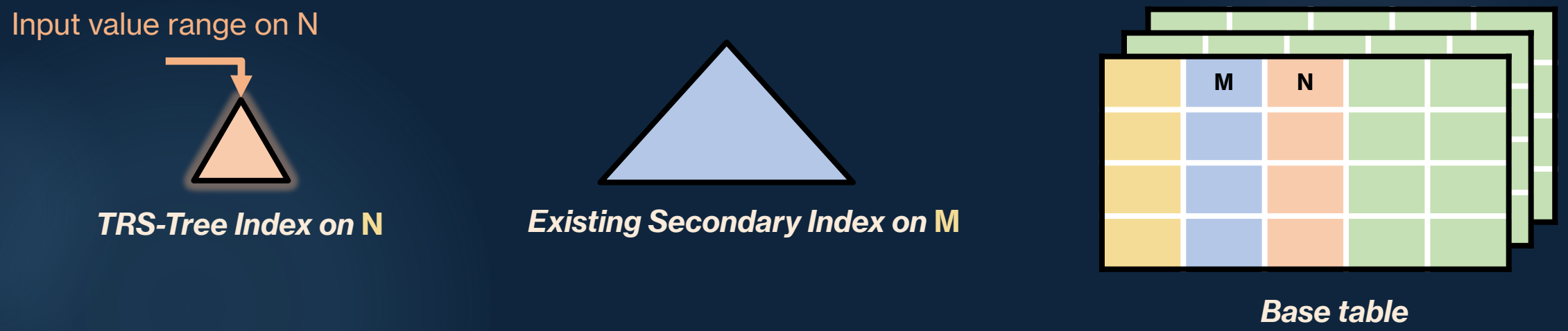


Existing Secondary Index on **M**

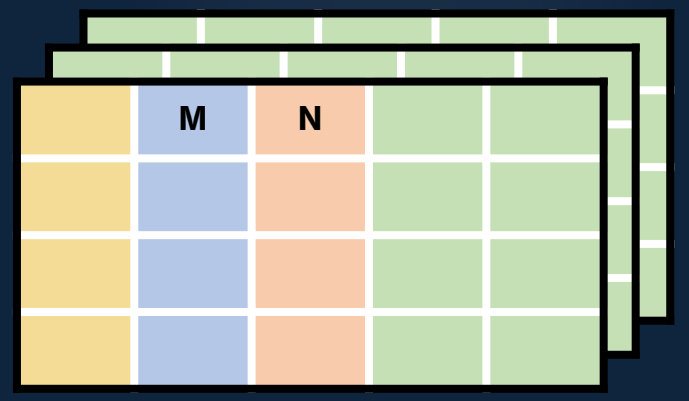


Base table

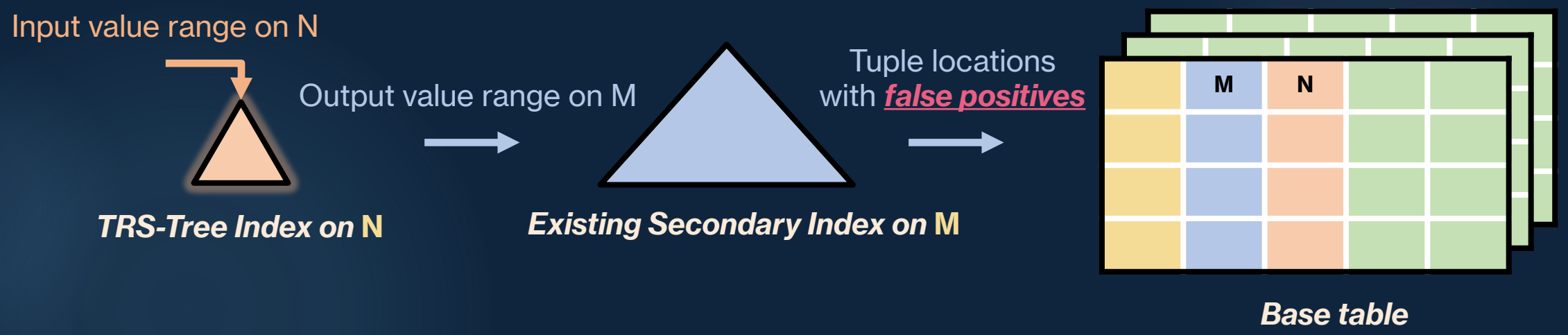
HERMIT: ACCESS METHOD



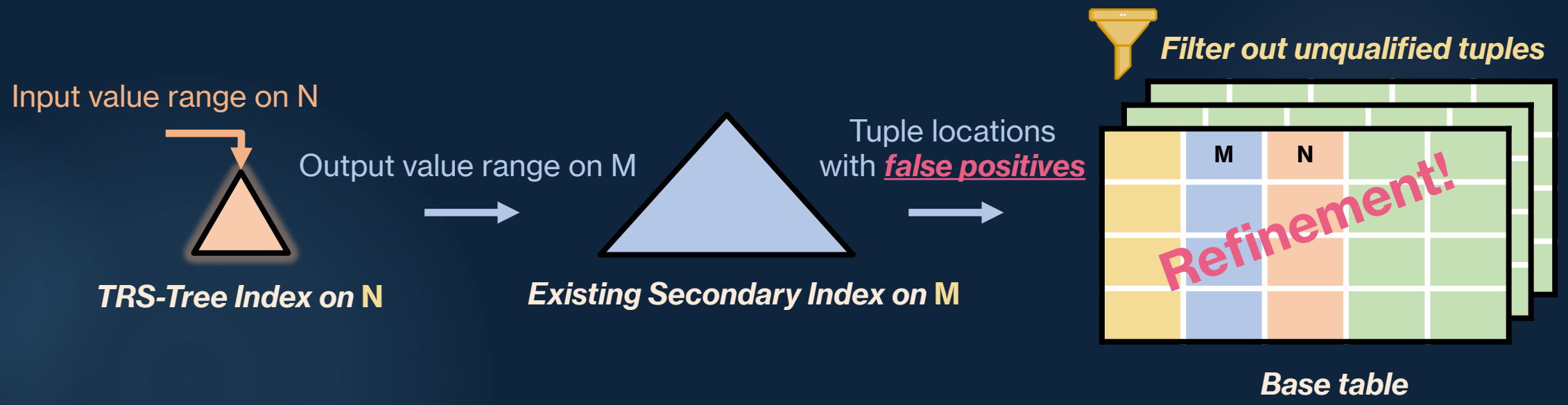
HERMIT: ACCESS METHOD



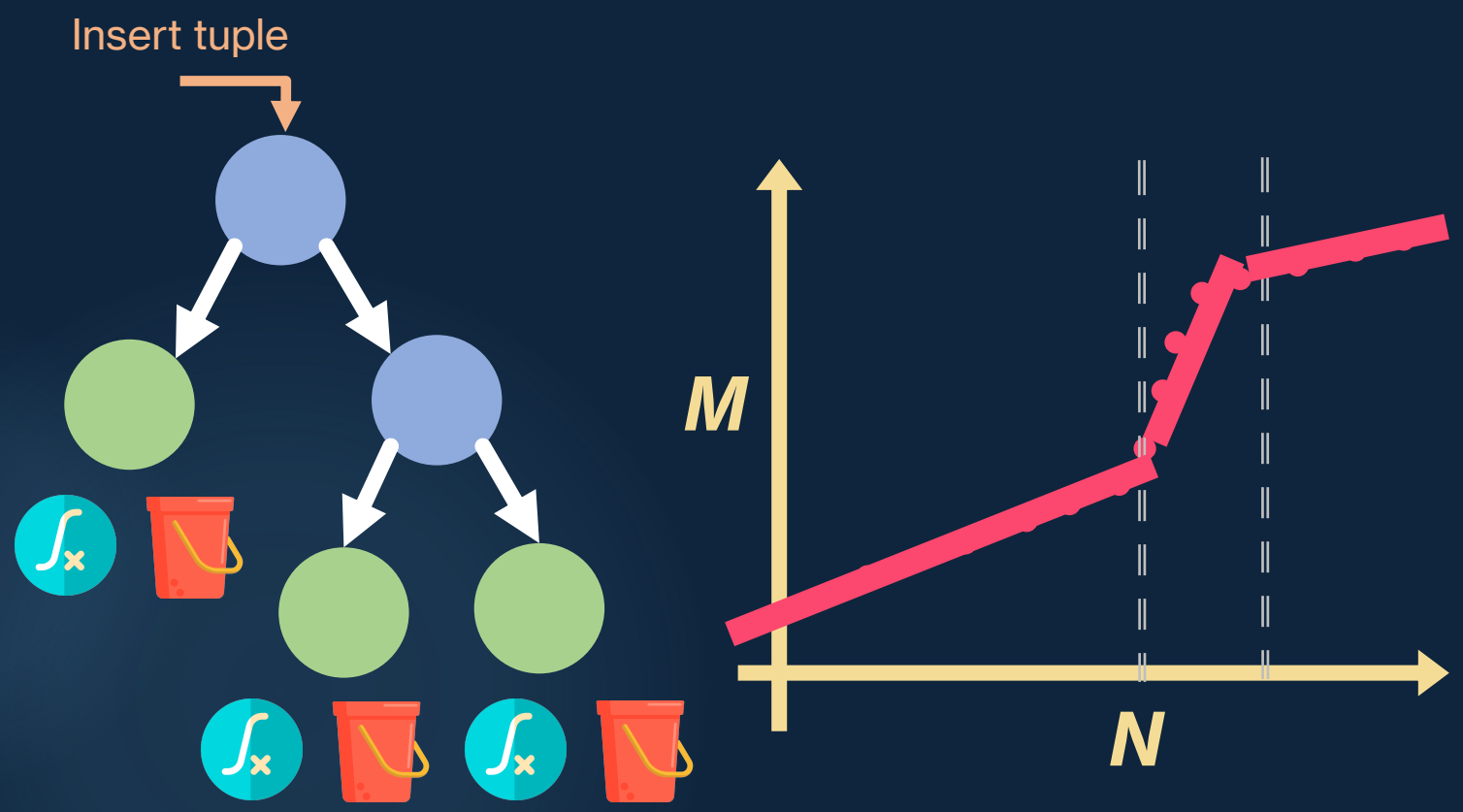
HERMIT: ACCESS METHOD



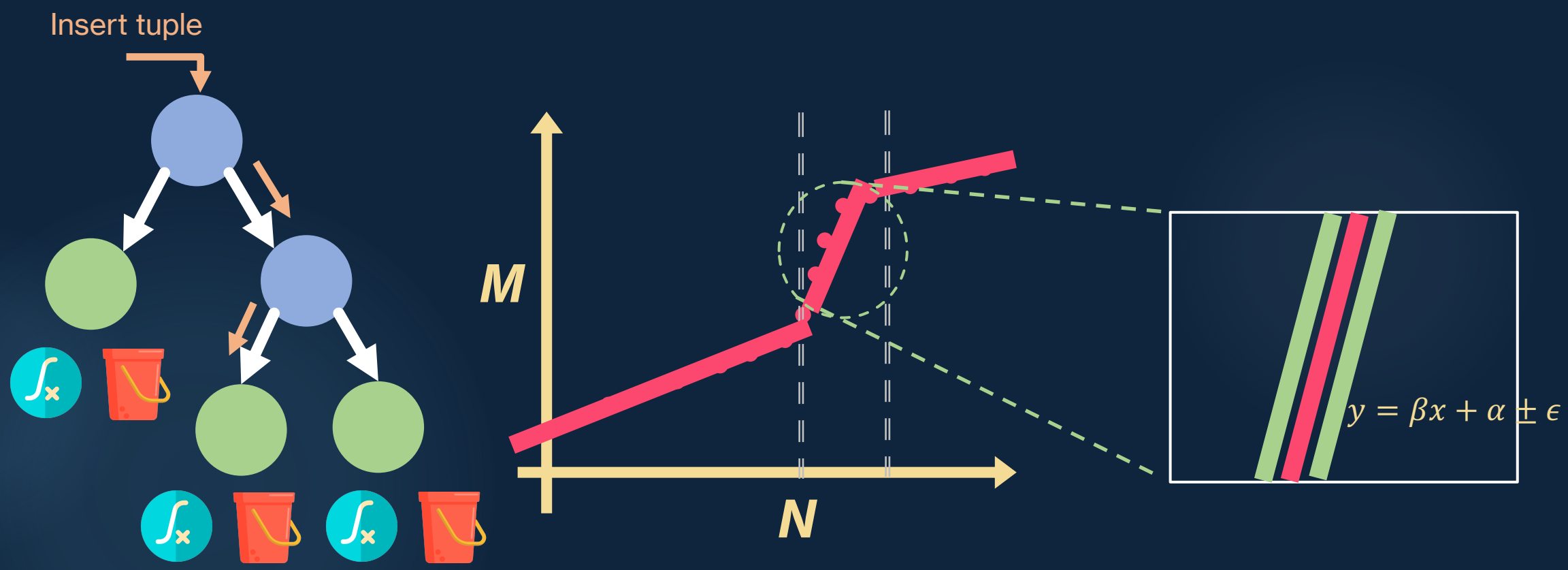
HERMIT: ACCESS METHOD



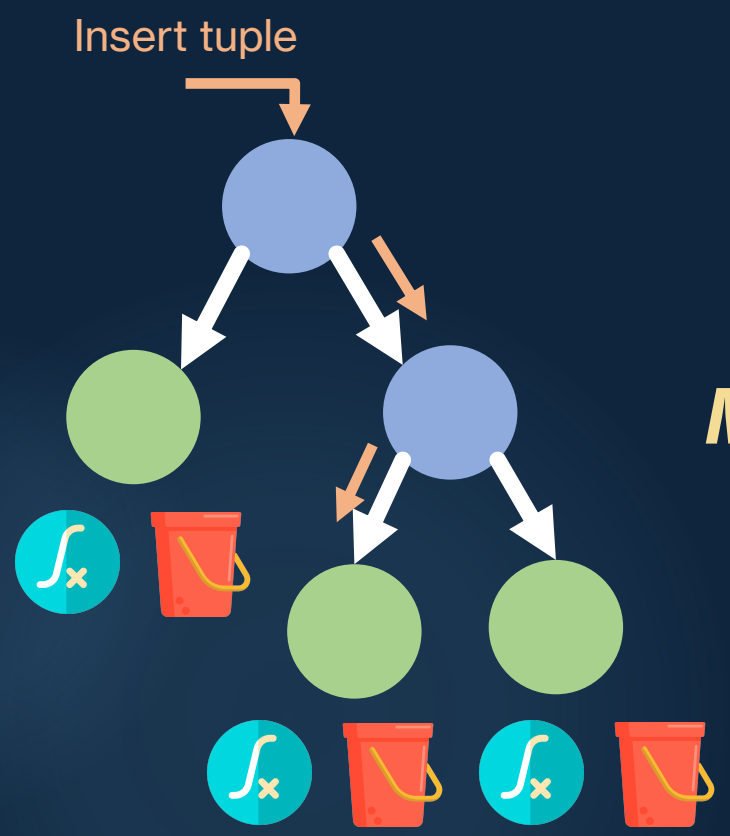
HERMIT: INSERT

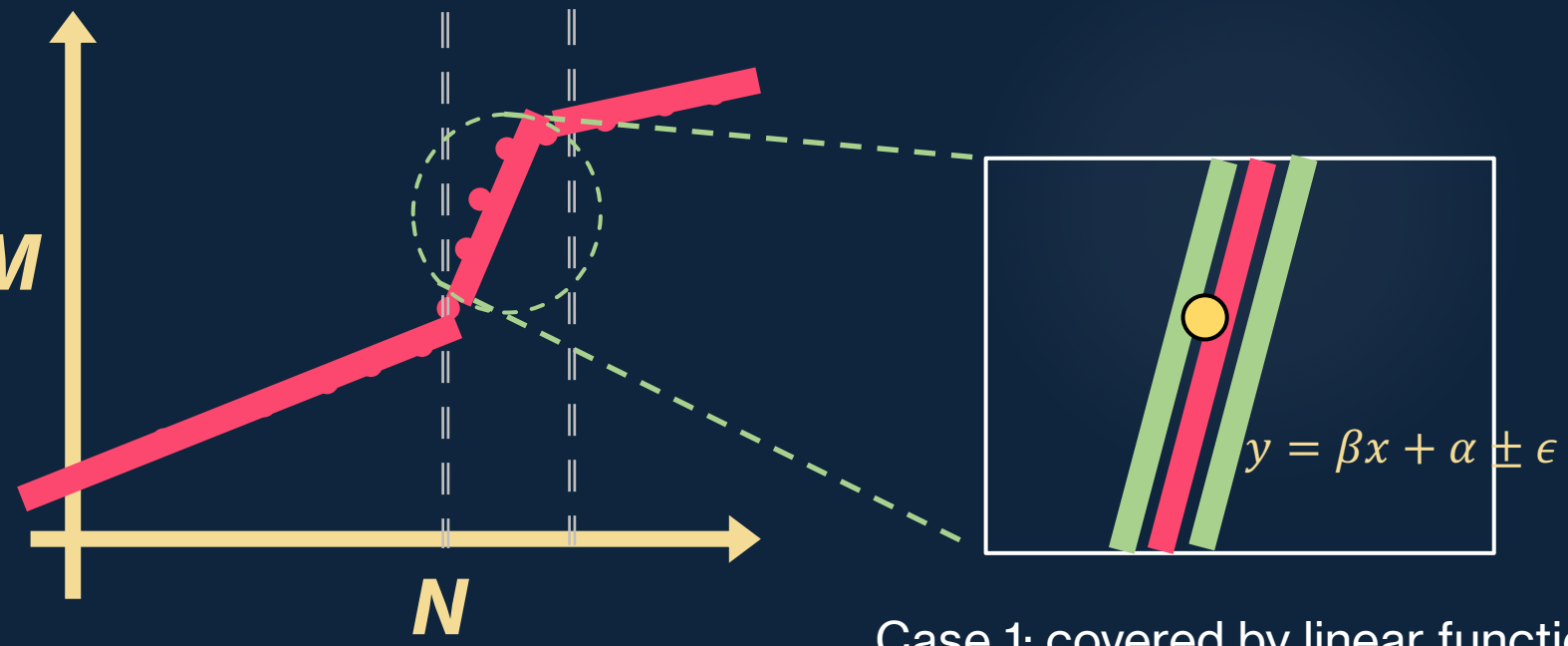


HERMIT: INSERT



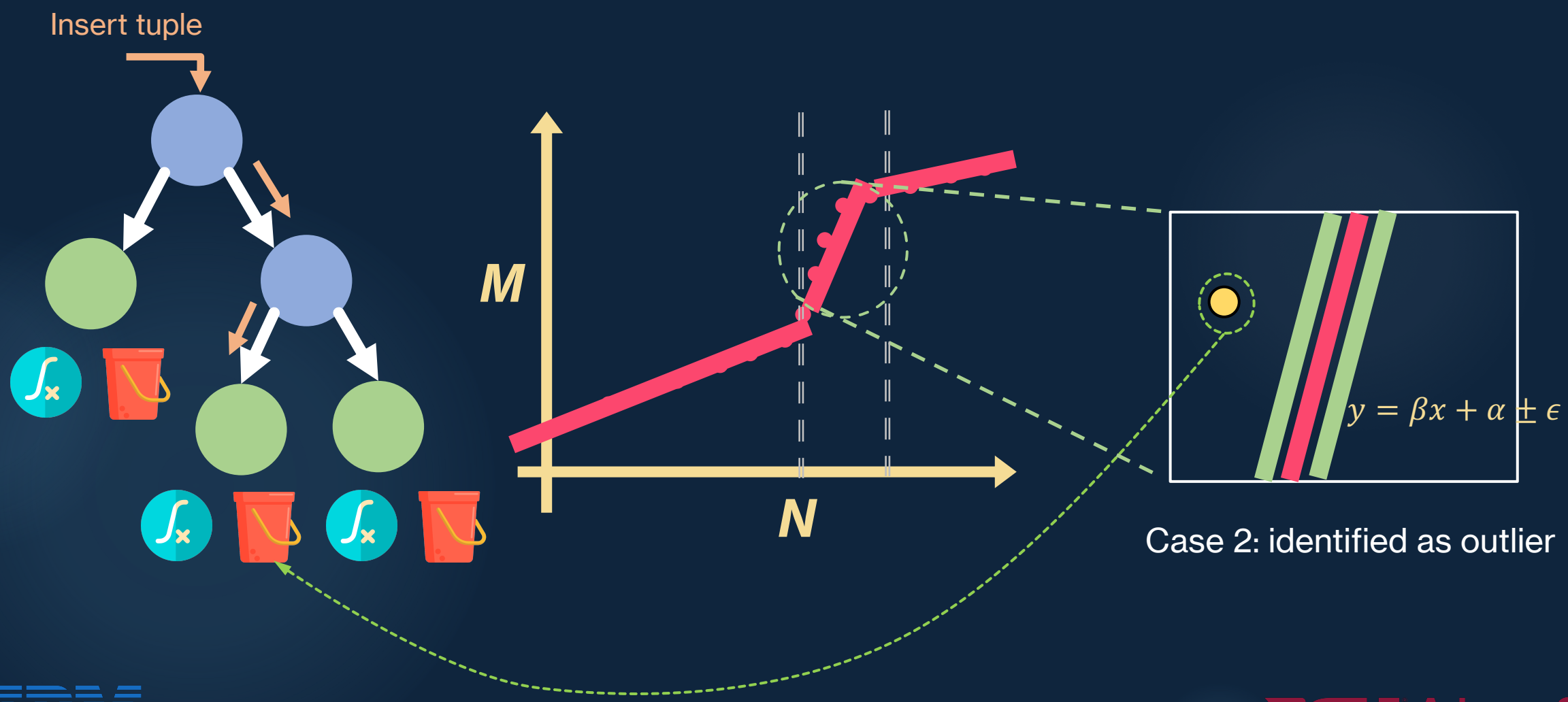
HERMIT: INSERT



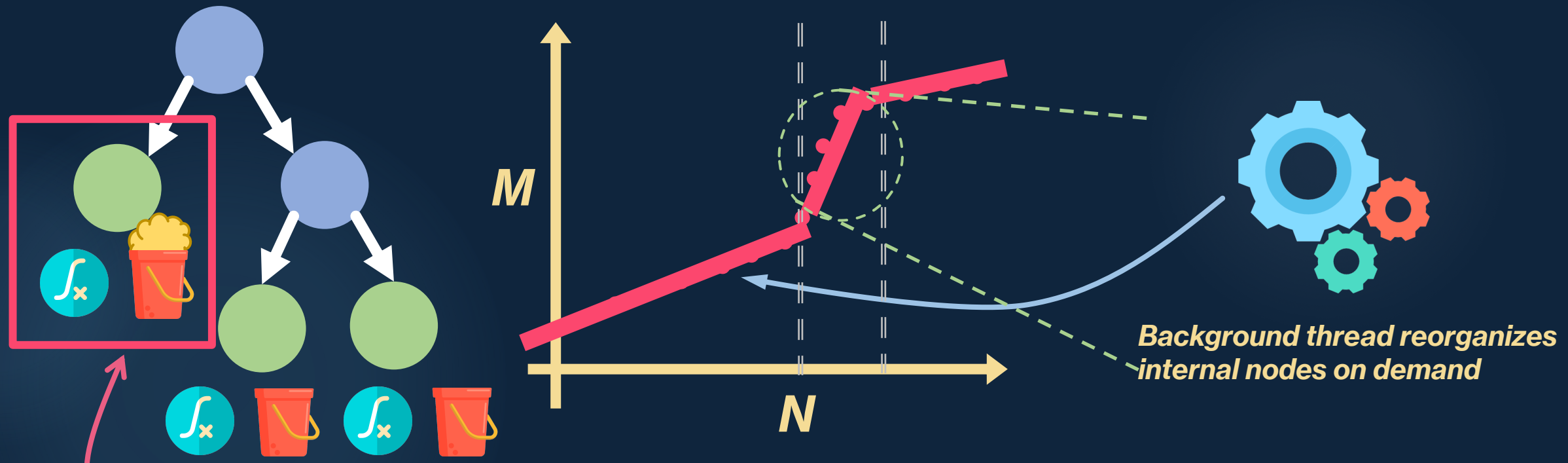


Case 1: covered by linear function

HERMIT: INSERT



HERMIT: MAINTENANCE



Once outlier bucket becomes full, then reorganization is required.

HERMIT: EVALUATION

- Benchmark
  - Stock: 100 stocks with highest and lowest prices
  - Sensor: 16 gas concentration sensors in different locations
  - Synthetic: Linear function, Sigmoid function
- Environment
  - In-memory performance
    - System: Main-memory database prototype
    - Baseline: B+-tree index

HERMIT: EVALUATION

- Benchmark
  - Stock: 100 stocks with highest and lowest prices
  - Sensor: 16 gas concentration sensors in different locations
  - ***Synthetic: Linear function, Sigmoid function***
- Environment
  - ***In-memory performance***
    - ***System: Main-memory database prototype***
    - ***Baseline: B+-tree index***

HERMIT: EVALUATION

Query: "select all the tuples with N's value falling between x and y"

- Benchmark
  - Stock: 100 stocks with highest and lowest prices
  - Sensor: 16 gas concentration sensors in different locations
  - ***Synthetic: Linear function, Sigmoid function***

$$M = \beta N + \alpha \qquad M = Sigmoid(N)$$

- Environment
  - In-memory performance
    - System: Main-memory database prototype
    - Baseline: B+-tree index

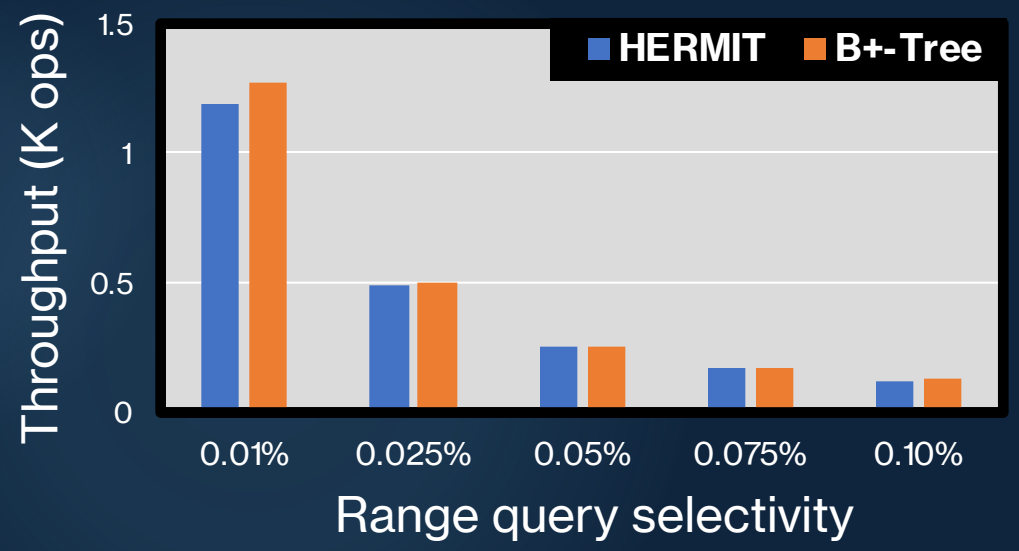
Key	M	N
1	100	115
2	200	204
3	150	165
4	400	410
...	...	...

v.s.

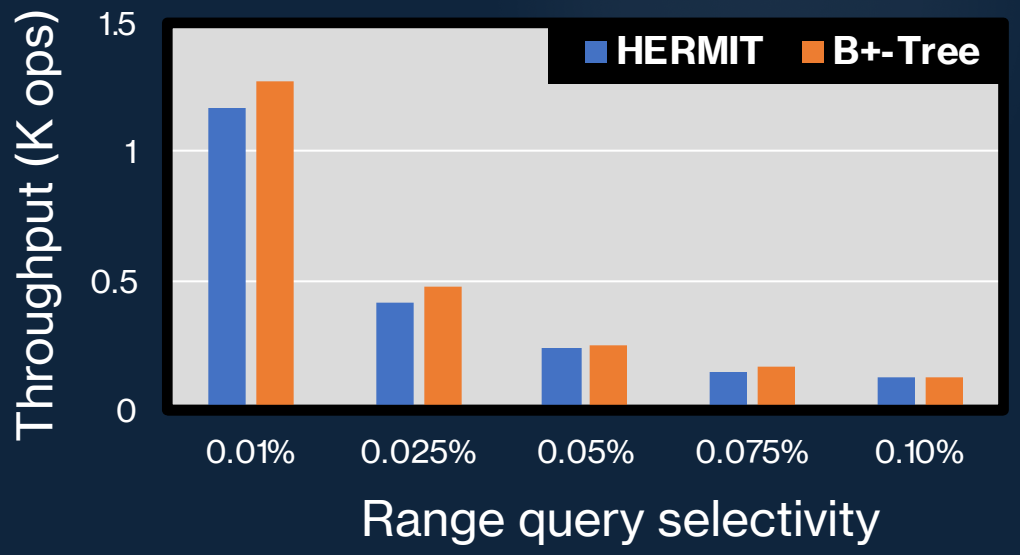
Hermit B+-Tree

HERMIT: EVALUATION

- Synthetic benchmark (range lookup throughput)



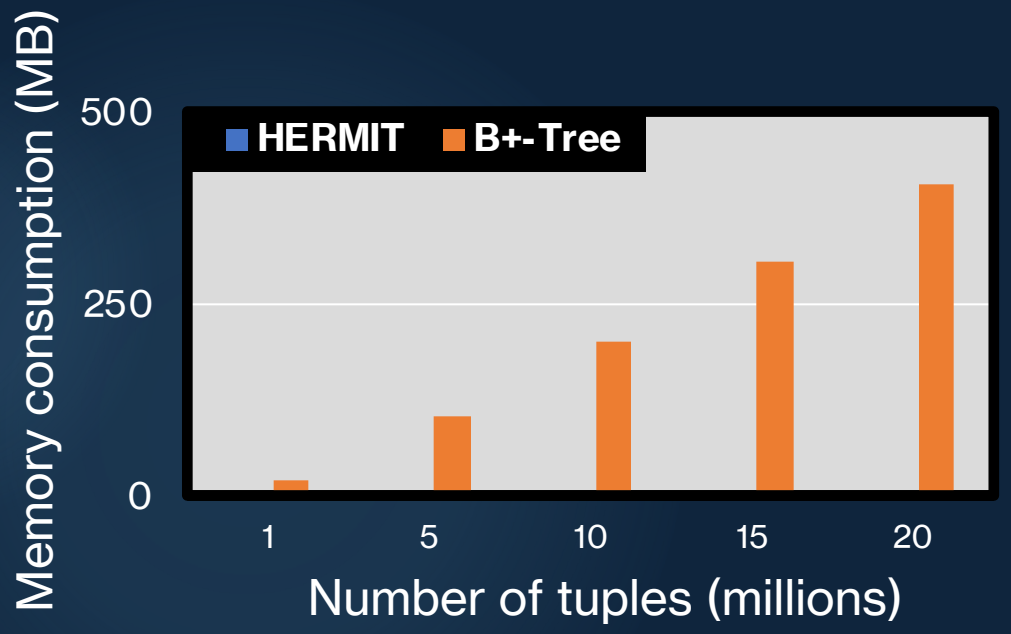
Linear function



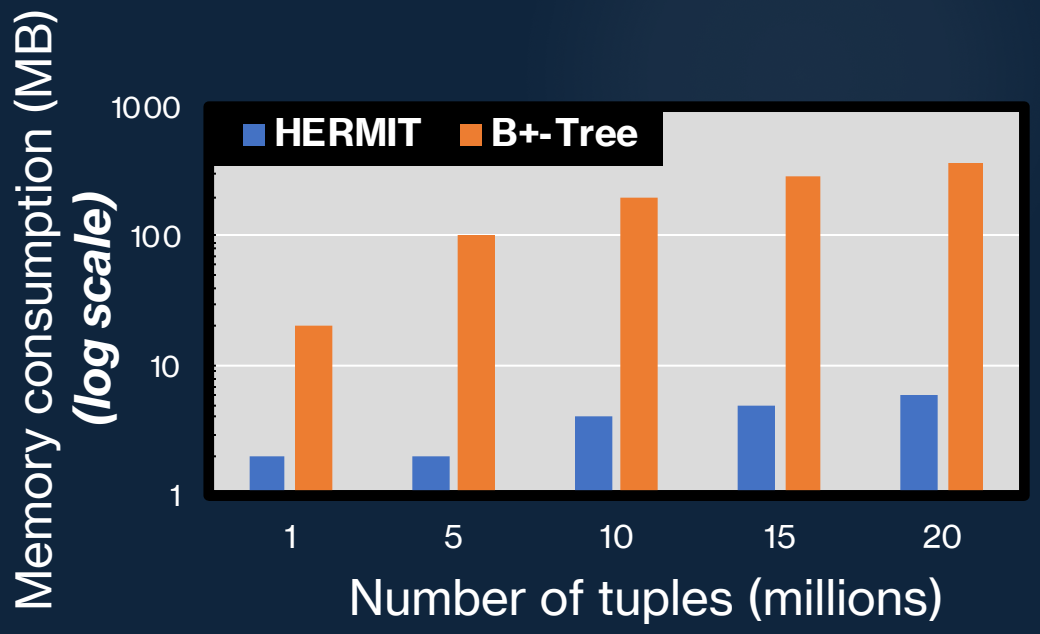
Sigmoid function

HERMIT: EVALUATION

- Synthetic benchmark (index memory consumption)



Linear function



Sigmoid function

HERMIT: CONCLUSION

- HERMIT is a succinct secondary indexing mechanism
  - Use ML-enhanced TRS tree to model correlation and capture outliers
  - Support all kinds of operations
  - Show good lookup / insertion performance and significantly reduce index size

HERMIT: EVALUATION

- Benchmark
  - Stock: 100 stocks with highest and lowest prices
  - Sensor: 16 gas concentration sensors in different locations
  - ***Synthetic: Linear function, Sigmoid function***
- Environment
  - ***In-memory performance***
    - ***System: Main-memory database prototype***
    - ***Baseline: B+-tree index***

HERMIT: EVALUATION

Query: "during which time periods do Stock X's highest price fall between Y and Z?"

- Benchmark
  - ***Stock: 100 stocks with highest and lowest prices***
  - Sensor: 16 gas concentration sensors in different locations
  - Synthetic: Linear function, Sigmoid function
- Environment
  - In-memory performance
    - System: Main-memory database prototype
    - Baseline: B+-tree index

Date	IBM (low)	IBM (high)	MSFT (low)	MSFT (high)	...	...
90/1/1	200	220	110	132		
90/1/2	180	190	120	180		
90/1/3	190	201	124	130		
90/1/4	195	207	116	125		
...	...	...	...	...		

HERMIT: EVALUATION

Query: "during which time periods do Stock X's highest price fall between Y and Z?"

- Benchmark
  - ***Stock: 100 stocks with highest and lowest prices***
  - Sensor: 16 gas concentration sensors in different locations
  - Synthetic: Linear function, Sigmoid function
- Environment
  - In-memory performance
    - System: Main-memory database prototype
    - Baseline: B+-tree index

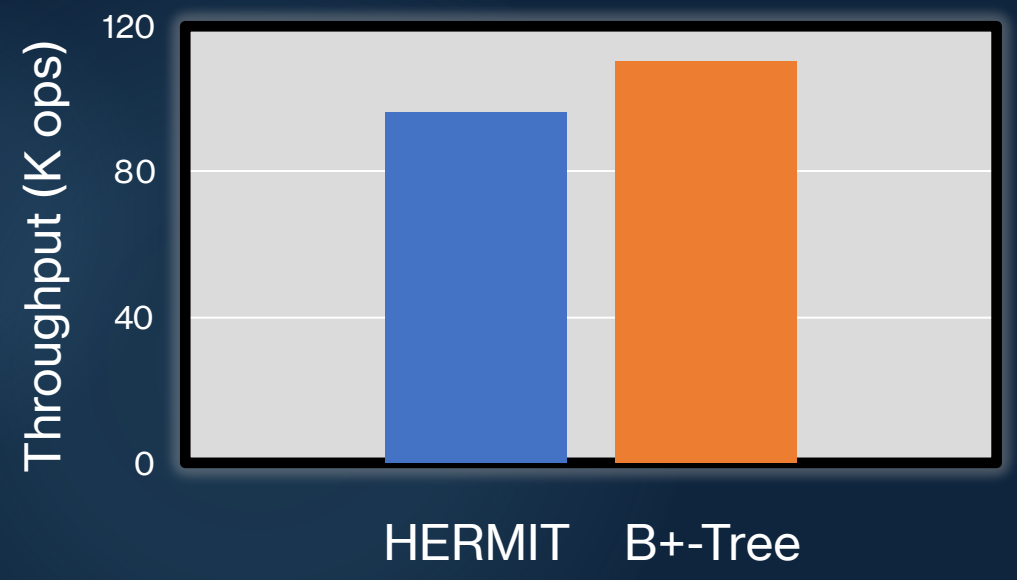
Date	IBM (low)	IBM (high)	MSFT (low)	MSFT (high)	...	...
90/1/1	200	220	110	132		
90/1/2	180	190	120	180		
90/1/3	190	201	124	130		
90/1/4	195	207	116	125		
...	...	...	...	...		

v.s. *v.s.*

Hermit B+-Tree

HERMIT: EVALUATION

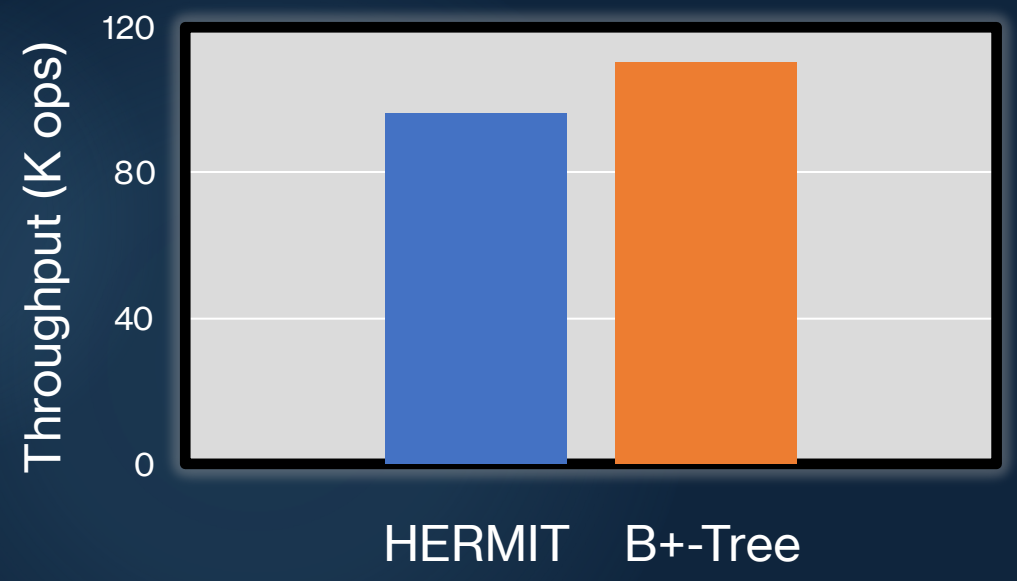
- Stock benchmark



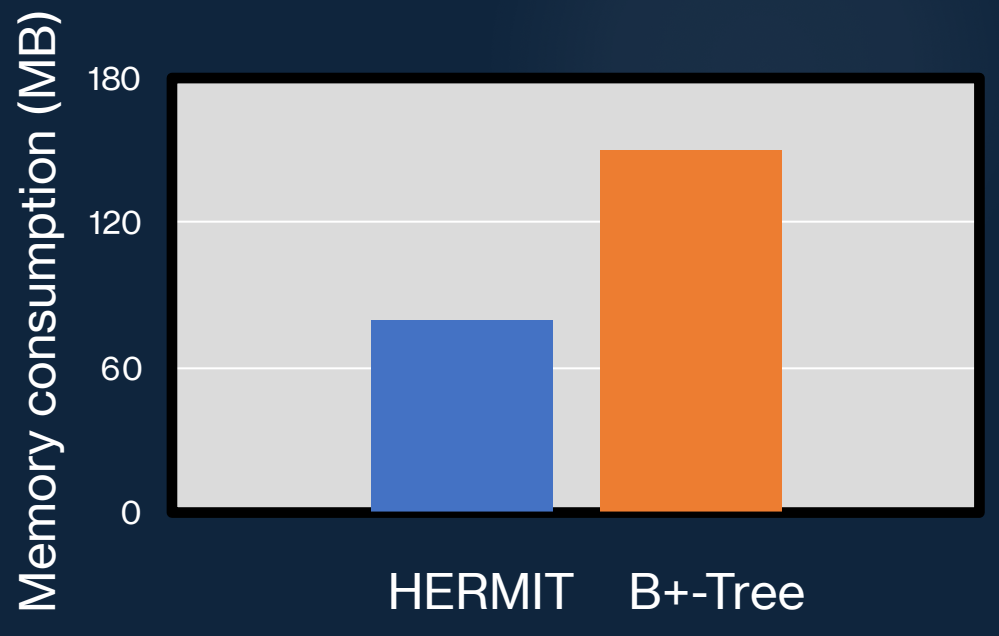
Range lookup throughput

HERMIT: EVALUATION

- Stock benchmark



Range lookup throughput



Overall memory consumption

HERMIT: COMPLEX MODEL

Training time for different ML models (scikit-learn)

Number of tuples	1 K	10 K	100 K
Linear regression	0.42 ms	0.81 ms	3.2 ms
SVR (RBF)	0.09 s	4.5 s	> 120 s
SVR (linear)	0.28 s	29 s	> 120 s
SVR (polynomial)	0.29 s	24 s	> 120 s