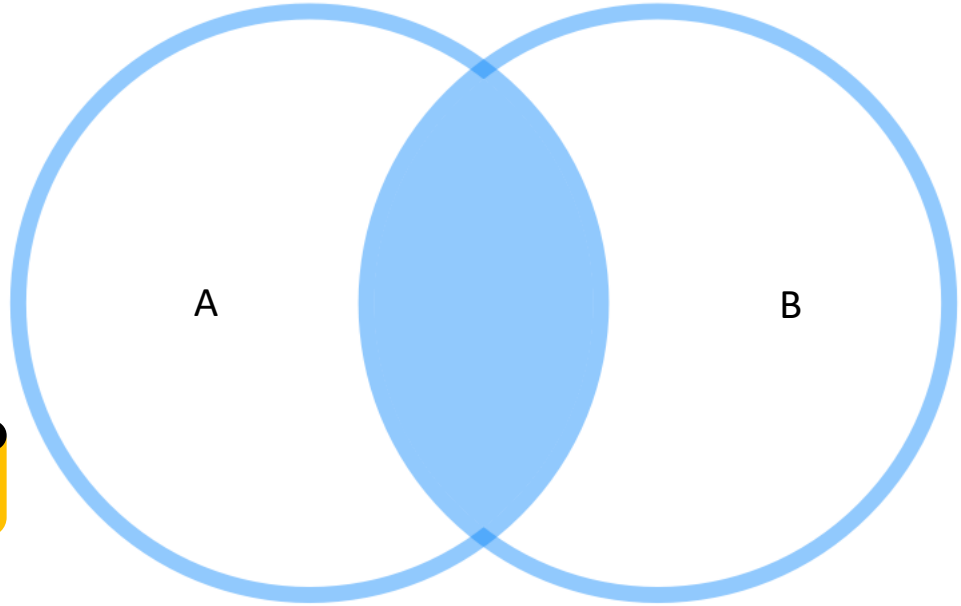
Inner joins and full outer joins

Merging: inner join



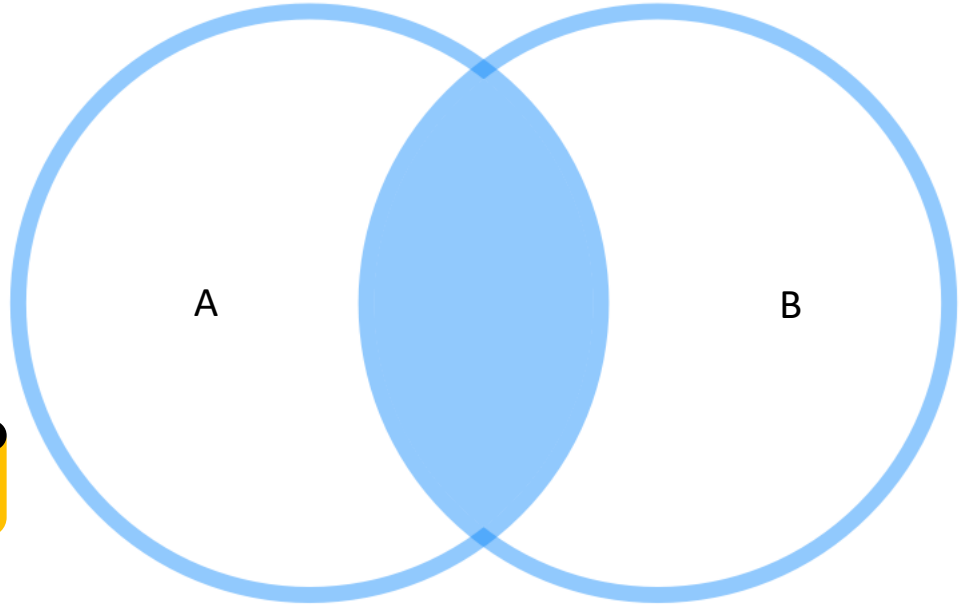
Inner joins return only the observations available in both datasets ❶

```
merge(A, B, by="ID", all=FALSE)
```

❷ Common identifier

❸ Only matching rows are returned

Merging: inner join



Inner joins return only the observations available in both datasets ❶

```
merge(A, B, by="ID", all=FALSE)
```

❷ Common identifier

❸ Only matching rows are returned

Inner join merges on common identifiers present in both data.tables (1/2)

A (myData)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	2005-11-15	1	199.95	107.00
172951	2008-08-29	1	199.95	108.00
120621	2007-10-19	1	99.95	49.00
149236	2005-11-14	1	39.95	18.95
149236	2007-12-06	1	79.95	35.00
…	…	…	…	…

B (CustData)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	1991-08-26	US-06332	2009-09-15
149332	m	1998-07-07	US-08873	2005-11-05
84374	m	1977-07-10	US-06400	1988-08-10
149236	f	1955-08-15	US-92646	1971-02-16
100001	m	1974-05-08	US-02332	1992-02-21
…	…	…	…	…

```
merge(myData, CustData, by="Customer")
```

Inner join merges on common identifiers present in both data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	2005-11-15	1	199.95	107.00
172951	2008-08-29	1	199.95	108.00
120621	2007-10-19	1	99.95	49.00
149236	2005-11-14	1	39.95	18.95
149236	2007-12-06	1	79.95	35.00
...	...	...	...	...

B (`CustData`)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	1991-08-26	US-06332	2009-09-15
149332	m	1998-07-07	US-08873	2005-11-05
84374	m	1977-07-10	US-06400	1988-08-10
149236	f	1955-08-15	US-92646	1971-02-16
100001	m	1974-05-08	US-02332	1992-02-21
...	...	...	...	...

Merge rows with the same customer ID if customer ID occurs in both tables

Customer	TransDate	PurchAmount	Cost	Gender	Birthdate	ZIP	JoinDate
149332	2005-11-15	1	199.95	m	1998-07-07	US-08873	05.11.2005
149236	2005-11-14	1	39.95	f	1955-08-15	US-92646	16.02.1971
149236	2001-06-12	1	79.95	f	1955-08-15	US-92646	16.02.1971
...	...	...	...	...	...	...	...

Inner join merges on common identifiers present in both data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	2005-11-15	1	199.95	107.00
172951	2008-08-29	1	199.95	108.00
120621	2007-10-19	1	99.95	49.00
149236	2005-11-14	1	39.95	18.95
149236	2007-12-06	1	79.95	35.00
...	...	...	...	...

B (`CustData`)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	1991-08-26	US-06332	2009-09-15
149332	m	1998-07-07	US-08873	2005-11-05
84374	m	1977-07-10	US-06400	1988-08-10
149236	f	1955-08-15	US-92646	1971-02-16
100001	m	1974-05-08	US-02332	1992-02-21
...	...	...	...	...

Merge rows with the same customer ID if customer ID occurs in both tables

Customer	TransDate	PurchAmount	Cost	Gender	Birthdate	ZIP	JoinDate
149332	2005-11-15	1	199.95	m	1998-07-07	US-08873	05.11.2005
149236	2005-11-14	1	39.95	f	1955-08-15	US-92646	16.02.1971
149236	2001-06-12	1	79.95	f	1955-08-15	US-92646	16.02.1971
...	...	...	...	...	...	...	...

Inner join merges on common identifiers present in both data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	2005-11-15	1	199.95	107.00
172951	2008-08-29	1	199.95	108.00
120621	2007-10-19	1	99.95	49.00
149236	2005-11-14	1	39.95	18.95
149236	2007-12-06	1	79.95	35.00
...	...	...	...	...

B (`CustData`)

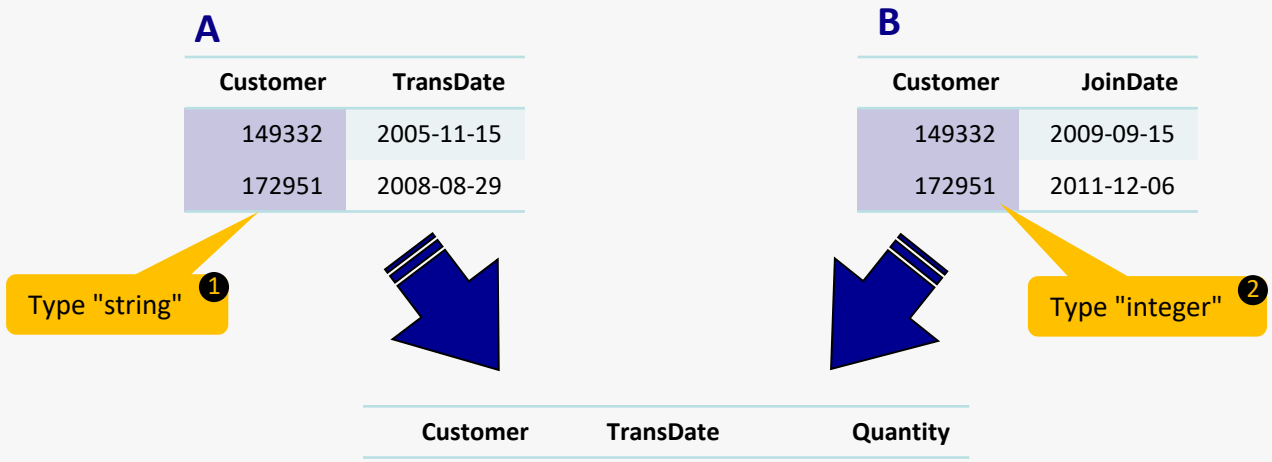
Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	1991-08-26	US-06332	2009-09-15
149332	m	1998-07-07	US-08873	2005-11-05
84374	m	1977-07-10	US-06400	1988-08-10
149236	f	1955-08-15	US-92646	1971-02-16
100001	m	1974-05-08	US-02332	1992-02-21
...	...	...	...	...

Merge rows with the same customer ID if customer ID occurs in both tables

Customer	TransDate	PurchAmount	Cost	Gender	Birthdate	ZIP	JoinDate
149332	2005-11-15	1	199.95	m	1998-07-07	US-08873	05.11.2005
149236	2005-11-14	1	39.95	f	1955-08-15	US-92646	16.02.1971
149236	2001-06-12	1	79.95	f	1955-08-15	US-92646	16.02.1971
...	...	...	...	...	...	...	...

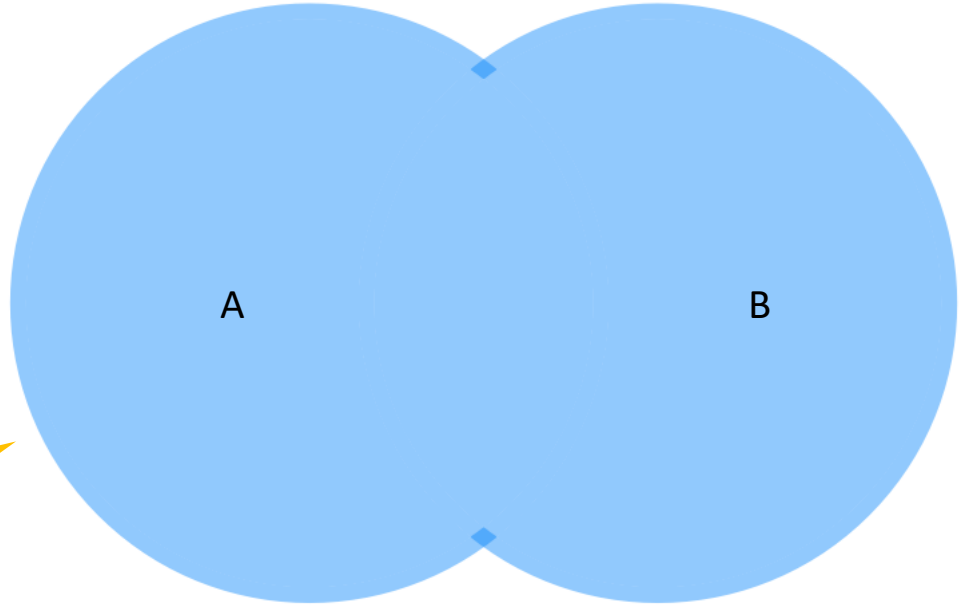
Sidenote: The common identifier needs to be the same class

Both customer columns need to be the same class. If one is an integer and the other is a string, the merge won't work.



- Returns an empty data.table

Merging: outer join



Outer joins return all the data available ❶

```
merge(A, B, by="ID", all=TRUE)
```

Common identifier ❷

Keep all rows from both datasets ❸

Full outer join merges on all common identifiers of both data.tables (1/2)

A (myData)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
149236	12.06.2001	1	79.95	35.00
…	…	…	…	…

B (CustData)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	26.08.1991	US-06332	15.09.2009
42886	f	04.05.1987	US-08055	12.06.2011
84374	m	10.07.1977	US-06400	10.08.1988
42291	m	12.07.1963	US-04533	23.07.1998
100001	m	08.05.1974	US-02332	21.02.1992
…	…	…	…	…

```
merge(myData, CustData, by="Customer", all=TRUE)
```

Full outer join merges on all common identifiers of both data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	2005-11-15	1	199.95	107.00
172951	2008-08-29	1	199.95	108.00
120621	2007-10-19	1	99.95	49.00
149236	2005-11-14	1	39.95	18.95
149236	2007-12-06	1	79.95	35.00
...	...	...	...	...

B (`CustData`)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	1991-08-26	US-06332	2009-09-15
149332	m	1998-07-07	US-08873	2005-11-05
84374	m	1977-07-10	US-06400	1988-08-10
149236	f	1955-08-15	US-92646	1971-02-16
100001	m	1974-05-08	US-02332	1992-02-21
...	...	...	...	...

Merge rows with the same customer IDs in both tables, otherwise fill entries of table with missing ID with NAs

Customer	TransDate	PurchAmount	Cost	Gender	DOB	ZIP	JoinDate
149332	2005-11-15	1	199.95	m	1998-07-07	US-08873	2005-11-05
149236	2005-11-14	1	39.95	f	1955-08-15	US-92646	1971-02-16
149236	2001-06-12	1	79.95	f	1955-08-15	US-92646	1971-02-16
172951	2008-08-29	1	199.95	NA	NA	NA	NA
120621	2007-10-19	1	99.95	NA	NA	NA	NA
80365	NA	NA	NA	f	1991-08-26	US-06332	2009-09-15
48374	NA	NA	NA	m	1977-07-10	US-06400	1988-08-10
...	...	...	...	...	...	...	...

Sidenote: Be careful when handling missing values (NA)

NA: Not available A missing value

Sidenote: Selecting missing and non-missing values in a data.table

- **Select missing values:**

```
myData[is.na(PurchAmount), ]
```

- **Wrong (does not yield any entries):**

```
myData[PurchAmount == NA, ]
```

- **Select non-missing values:**

```
myData[!is.na(PurchAmount), ]
```

Exclamation mark before `is.na()` (1)

- **Select rows with no missing values:**

```
myData[complete.cases(myData), ]
```

Be careful: incomplete case is dropped from the data (2)

Sidenote: Selecting missing and non-missing values in a data.table

- **Select missing values:**

```
myData[is.na(PurchAmount), ]
```

- **Wrong (does not yield any entries):**

```
myData[PurchAmount == NA, ]
```

- **Select non-missing values:**

```
myData[!is.na(PurchAmount), ]
```

Exclamation mark before `is.na()` (1)

- **Select rows with no missing values:**

```
myData[complete.cases(myData), ]
```

Be careful: incomplete case is dropped from the data (2)

Sidenote: Selecting missing and non-missing values in a data.table

- **Select missing values:**

```
myData[is.na(PurchAmount), ]
```

- **Wrong (does not yield any entries):**

```
myData[PurchAmount == NA, ]
```

- **Select non-missing values:**

```
myData[!is.na(PurchAmount), ]
```

Exclamation mark before `is.na()` (1)

- **Select rows with no missing values:**

```
myData[complete.cases(myData), ]
```

Be careful: incomplete case is dropped from the data (2)

Sidenote: Aggregating with missing values

```
> mean(c(1 ,2, NA, 3))
 NA
> mean(c(1, 2, NA, 3), na.rm = TRUE)
 2
```

NA remove = TRUE (1)

```
myData[, sum(PurchAmount, na.rm = TRUE), by = Customer]
```

NA remove = TRUE (2)

Sidenote: Aggregating with missing values

```
> mean(c(1 ,2, NA, 3))
 NA
> mean(c(1, 2, NA, 3), na.rm = TRUE)
 2
```

NA remove = TRUE (1)

```
myData[, sum(PurchAmount, na.rm = TRUE), by = Customer]
```

NA remove = TRUE (2)

Sidenote: Aggregating with missing values

```
> mean(c(1 ,2, NA, 3))
 NA
> mean(c(1, 2, NA, 3), na.rm = TRUE)
 2
```

NA remove = TRUE 1

```
myData[, sum(PurchAmount, na.rm = TRUE), by = Customer]
```

NA remove = TRUE 2

Sidenote: Aggregating with missing values

```
> mean(c(1 ,2, NA, 3))
 NA
> mean(c(1, 2, NA, 3), na.rm = TRUE)
 2
```

NA remove = TRUE (1)

```
myData[, sum(PurchAmount, na.rm = TRUE), by = Customer]
```

NA remove = TRUE (2)

Sidenote: Merging with missing values

A

ID	Age
NA	13
NA	25
1	31
2	40

B

ID	Gender
NA	f
1	m
2	f
NA	f

```
merge(A, B, by = "ID", all = FALSE)
```

ID	Age	Gender
NA	13	f
NA	13	f
NA	25	f
NA	25	f
1	31	m
2	40	f

If information for keys are missing, each will be treated as an individual case

Build the cartesian products of data.tables (1/2)

A (myData)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
80365	24.02.2006	1	99.95	108.00
…	…	…	…	…

B (CustData)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	26.08.1991	US-06332	15.09.2009
149332	m	07.07.1998	US-08873	05.11.2005
84374	m	10.07.1977	US-06400	10.08.1988
149236	f	15.08.1955	US-92646	16.02.1971
80365	f	26.08.1991	US-06349	15.09.2009
…	…	…	…	…

1 A common key column is needed

```
merge(A, B, by=""Customer", all=TRUE, allow.cartesian=TRUE)
```

2 Used if you need all possible combinations

Build the cartesian products of data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
80365	24.02.2006	1	99.95	108.00
...	...	...	...	...

B (`CustData`)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	26.08.1991	US-06332	15.09.2009
149332	m	07.07.1998	US-08873	05.11.2005
84374	m	10.07.1977	US-06400	10.08.1988
149236	f	15.08.1955	US-92646	16.02.1971
80365	f	26.08.1991	US-06349	15.09.2009
...	...	...	...	...

Copy row from Table A if there are multiple matches in Table B.

Customer	TransDate	PurchAmount	Cost	Gender	DOB	ZIP	JoinDate
149332	15.11.2005	1	199.95	m	07.07.1998	US-08873	05.11.2005
....	....	...	...	...	...	...	...
120621	19.10.2007	1	99.95	NA	NA	NA	NA
80365	24.02.2006	1	99.95	f	26.08.1991	US-06332	15.09.2009
80365	24.02.2006	1	99.95	f	26.08.1991	US-06349	15.09.2009

Build the cartesian products of data.tables (2/2)

A (`myData`)

Customer	TransDate	Quantity	PurchAmount	Cost
149332	15.11.2005	1	199.95	107.00
172951	29.08.2008	1	199.95	108.00
120621	19.10.2007	1	99.95	49.00
149236	14.11.2005	1	39.95	18.95
80365	24.02.2006	1	99.95	108.00
...	...	...	...	...

B (`CustData`)

Customer	Gender	Birthdate	ZIP	JoinDate
80365	f	26.08.1991	US-06332	15.09.2009
149332	m	07.07.1998	US-08873	05.11.2005
84374	m	10.07.1977	US-06400	10.08.1988
149236	f	15.08.1955	US-92646	16.02.1971
80365	f	26.08.1991	US-06349	15.09.2009
...	...	...	...	...

Copy row from Table A if there are multiple matches in Table B.

Customer	TransDate	PurchAmount	Cost	Gender	DOB	ZIP	JoinDate
149332	15.11.2005	1	199.95	m	07.07.1998	US-08873	05.11.2005
....	....	...	...	...	...	...	...
120621	19.10.2007	1	99.95	NA	NA	NA	NA
80365	24.02.2006	1	99.95	f	26.08.1991	US-06332	15.09.2009
80365	24.02.2006	1	99.95	f	26.08.1991	US-06349	15.09.2009