A *very* quick introduction to dplyr

Sean C. Anderson
sean@seananderson.ca
December 2013

Why dplyr?

- *much* faster
- specializes in data.frame output
- works with many data inputs
- flexible and powerful grammar
- built for big(ger) data
- useful print statements

Why not dplyr?

- not available in compiled version yet
- under rapid development (it may change)
- it's very new (i.e. I don't know it well)
- for most purposes, plyr will be fine

```
tbl_df()
tbl_dt()
...
```

```
hflights_df <- tbl_df(hflights)
hflights_df
```

See `dplyr/vignettes/introduction.Rmd`

```
Source: local data frame [227,496 x 21]

     Year Month DayofMonth DayOfWeek DepTime ArrTime
5424 2011     1          1         6    1400    1500
5425 2011     1          2         7    1401    1501
5426 2011     1          3         1    1352    1502
5427 2011     1          4         2    1403    1513
5428 2011     1          5         3    1405    1507
5429 2011     1          6         4    1359    1503
5430 2011     1          7         5    1359    1509
..    ...   ...        ...       ...     ...     ...
Variables not shown: UniqueCarrier (chr), FlightNum (int),
  TailNum (chr), ActualElapsedTime (int), AirTime (int),
  ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr),
  Distance (int), TaxiIn (int), TaxiOut (int), Cancelled
  (int), CancellationCode (chr), Diverted (int)
```

select()

arrange()

filter()

```
filter(hflights_df, Month == 1,
  DayofMonth == 1)
```

```
select(hflights_df, Year, Month, DayOfWeek)
select(hflights_df, Year:DayOfWeek)
select(hflights_df, -(Year:DayOfWeek))
```

group_by()

summarise()

do()