# Fancy vectors

There are three kinds of vectors whose proxy has a different size than the actual vector value. Compressed vectors, and raw vectors.

There are also vectors whose size matches, but which require overriding some primitive vctrs operations such as equality.

## Compressed vectors

These vectors have in common that they wrap another vector type with a vector of indices to map the stored values from compressed space to value space.

* RLE vectors
* vec\_group\_rle(mtcars[c("cyl", "am")])  
  #> <vctrs\_group\_rle[17][n = 6]>  
  #> [1] 1x2 2x1 3x1 4x1 3x1 4x1 5x2 3x2 4x6 2x3 5x1 4x4 2x3 6x1 1x1  
  #> [16] 6x1 2x1
* Sparse vectors, e.g.
* x <- Matrix::sparseVector(c(2, 2.5, 3), c(2, 5, 9), 10)  
  unclass(x)  
  #> <S4 Type Object>  
  #> attr(,"x")  
  #> [1] 2.0 2.5 3.0  
  #> attr(,"length")  
  #> [1] 10  
  #> attr(,"i")  
  #> [1] 2 5 9

## Raw vectors

Unlike compressed vectors, raw vectors are an actual value array. However R hard-code the element size to 8 bits. To make raw vectors useful, we need a way of specifying the bit width of an element.

One example of this would be an int64\_t vector wrapped in a RAWSXP. The vector has contiguous values but its R length() is 8 times larger than the number of elements in the C array.

Having general support for raw vectors might be a good alternative to ALTREP for the cases where the data is not generated on the spot. Advantages:

* We control the implementation in package space
* Potentially easier to specify
* Available on old R versions

We’d still rely on ALTREP for generated and lazy data.

## Geospatial vectors

They don’t lend themselves well to memory proxies because the computations to determine geospatial equality are complex.

# Customisation points

## Size and slice

For compressed/expanded vectors, we could either provide a general customisation point for mapping indices from value space to proxy space. However it seems easier to hard-code the different kinds of fancy proxies and special-case them in vctrs.

## Equality and hashing

For compressed vectors, memory equality and hashing is sufficient.

For geospatial equality and raw equality, we need a way to override the equality and hashing operation.

## Ordering

Technically the vec\_proxy\_compare() generic could take care of it. It is possible allowing overriding the comparison operator could be useful for performance though.

## Internal representation

Compressed and raw vectors can’t be represented by their data because their size wouldn’t match the size of data frames when they are included as columns.

For this reason, we need a dummy vector of the actual size.

* The dummy vector could be implemented with ALTREP on recent R. This would make it free in terms of memory and computation.
* The actual vector data would be contained in attributes. All vctrs primitive operations (slice, concatenation, etc) would need to be forwarded to the data.

## Missing values

For compressed vectors, the missing values are encoded natively.

For raw vectors, there are two options.

1. The proxy could expose an external vector of locations for missing values, arrow-style. This vector could possibly be implemented as a sparse vector, and will need to be part of all operations (slicing, concatenation, …).
2. Go with R-style missing values and just allow overriding missingness detection, the same way we’d allow overriding equality / ordering.
3. Use the dummy vector (see *Internal representation* section). Missing values inside the dummy vector could represent missingness.