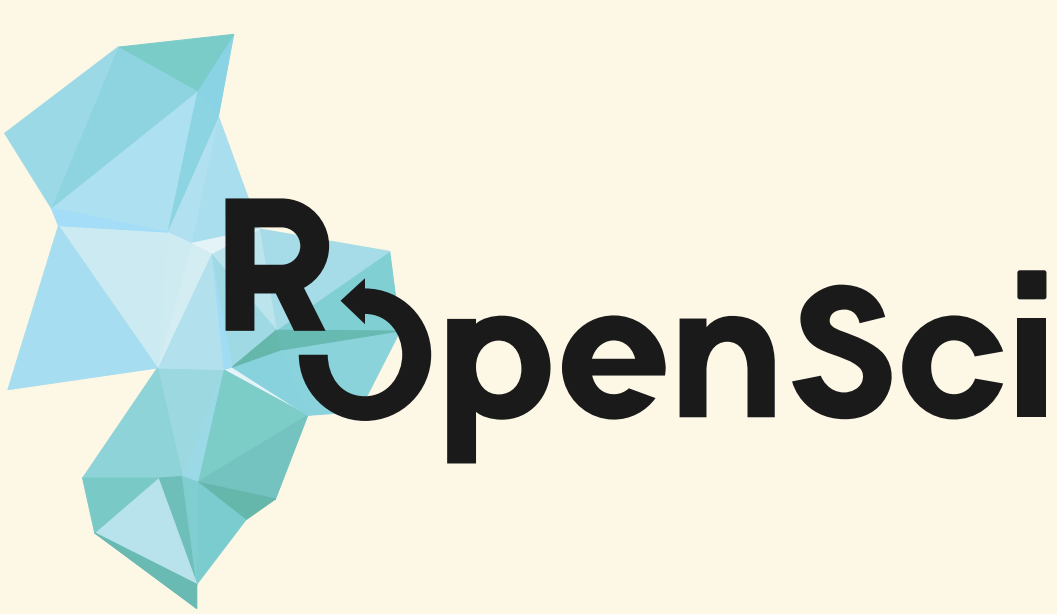
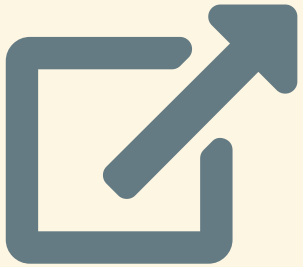
taxonomic data methods for R & Python

Scott Chamberlain (@sckottie) and Zachary Foster

UC Berkeley / rOpenSci





scotttalks.info/bosc18

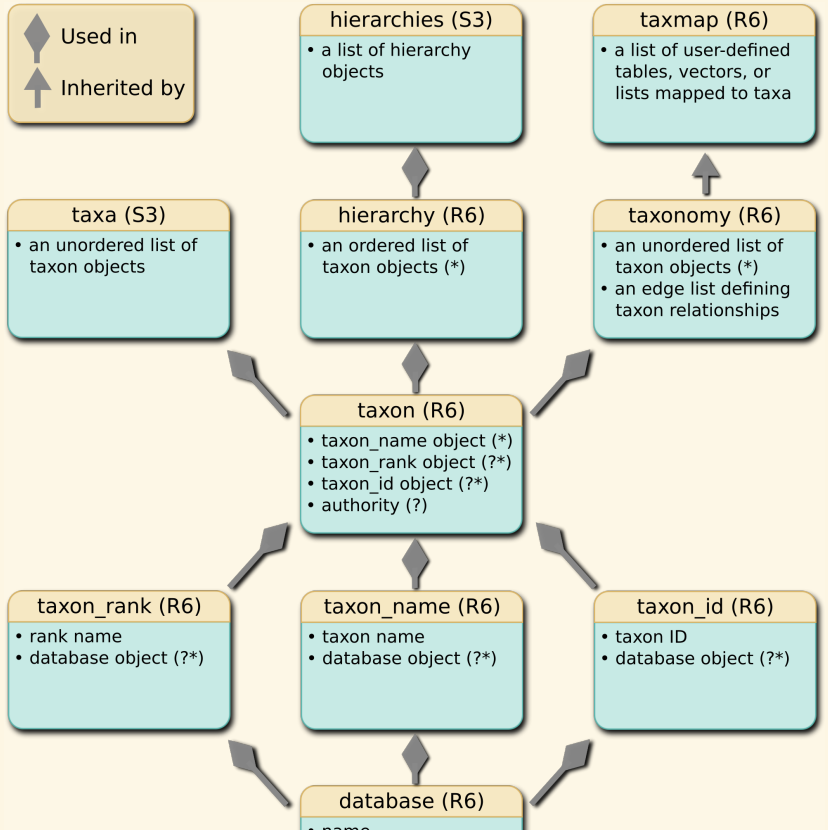
challenges of taxonomic data

- taxonomic data is hierarchical
- "taxa" can be names, classifications of names, or ids
- diff. sources of taxonomic data (e.g. NCBI vs. COL)
  - may have different names and ids for the same taxon
- often associated with other data
  - ideally taxa filtering is linked to data

taxa: the R package

github.com/ropensci/taxa

- classes to hold taxa, taxonomies, and associated data
- flexible parsers to convert raw data to those classes
- dplyr-inspired functions to easily manipulate classes
- any filtering/subsetting keeps all data linked together
- flexible base for other packages to build on
  - R: taxize, taxizedb, metacoder, pytaxize, etc.
  - Python: pytaxize, etc.



taxmap: user defined data mapped to taxonomy

```
<Taxmap>
  17 taxa: b. Mammalia, c. Plantae, d. Felidae ... p. sapiens, q. lycopersicum, r. tuberosum
  17 edges: NA->b, NA->c, b->d, b->e, b->f, c->g, d->h, d->i ... g->l, h->m, i->n, j->o, k->p, l->q, l->r
  4 data sets:
    info:
      # A tibble: 6 x 4
        taxon_id name  n_legs dangerous
        <chr>    <chr>  <dbl> <lgl>
      1 m        tiger      4 TRUE
      2 n        cat        4 FALSE
      3 o        mole       4 FALSE
      # ... with 3 more rows
    phylopic_ids: a named vector of 'character' with 6 items
       m. e148eabb-f138-43c6-b1e4-5cda2180485a ... r. 63604565-0406-460b-8cb8-1abe954b3f3a
    foods: a list of 6 items named by taxa:
       m, n, o, p, q, r
    abund:
      # A tibble: 8 x 5
        taxon_id code  sample_id count taxon_index
        <chr>    <fct> <fct>     <dbl>       <int>
      1 m        T     A             1           1
      2 n        C     A             2           2
      3 o        M     B             5           3
      # ... with 5 more rows
  1 functions:
    reaction
```

taxmap: user defined data mapped to taxonomy

```
<Taxmap>
  17 taxa: b, Mammalia, c, Plantae, d, Felidae ... p, sapiens, q, lycopersicum, r, tuberosum
  17 edges: NA->b, NA->c, b->d, b->e, b->f, c->g, d->h, d->i ... g->l, h->m, i->n, j->o, k->p, l->q, l->r
  4 data sets:
    info:
      # A tibble: 6 x 4
        taxon_id name  n_legs dangerous
        <chr>    <chr>  <dbl> <lgl>
      1 m        tiger      4 TRUE
      2 n        cat        4 FALSE
      3 o        mole       4 FALSE
      # ... with 3 more rows
    phylopic_ids: a named vector of 'character' with 6 items
       m, e148eabb-f138-43c6-b1e4-5cda2180485a ... r, 63604565-0406-460b-8cb8-1abe954b3f3a
    foods: a list of 6 items named by taxa:
       m, n, o, p, q, r
    abund:
      # A tibble: 8 x 5
        taxon_id code  sample_id count taxon_index
        <chr>    <fct> <fct>     <dbl>       <int>
      1 m        T     A             1           1
      2 n        C     A             2           2
      3 o        M     B             5           3
      # ... with 5 more rows
  1 functions:
    reaction
```

taxonomy condensed to edges and nodes

taxmap: user defined data mapped to taxonomy

```
<Taxmap>
  17 taxa: b. Mammalia, c. Plantae, d. Felidae ... p. sapiens, q. lycopersicum, r. tuberosum
  17 edges: NA->b, NA->c, b->d, b->e, b->f, c->g, d->h, d->i ... g->l, h->m, i->n, j->o, k->p, l->q, l->r
  4 data sets:
    info:
      # A tibble: 6 x 4
        taxon_id name  n_legs dangerous
        <chr>    <chr>  <dbl> <lgl>
      1 m        tiger      4 TRUE
      2 n        cat        4 FALSE
      3 o        mole       4 FALSE
      # ... with 3 more rows
    phylopic_ids: a named vector of 'character' with 6 items
       m. e148eabb-f138-43c6-b1e4-5cda2180485a ... r. 63604565-0406-460b-8cb8-1abe954b3f3a
    foods: a list of 6 items named by taxa:
       m, n, o, p, q, r
    abund:
      # A tibble: 8 x 5
        taxon_id code  sample_id count taxon_index
        <chr>    <fct> <fct>     <dbl>       <int>
      1 m        T     A             1           1
      2 n        C     A             2           2
      3 o        M     B             5           3
      # ... with 5 more rows
  1 functions:
    reaction
```

Arbitrary data linked to taxa

manipulate taxonomic data

Subset taxonomy and data to one taxon:

```
filter_taxa(x, taxon_names == "Plantae", subtaxa = TRUE)
```

Subset taxonomy to one rank:

```
filter_taxa(x, taxon_ranks == "genus", supertaxa = TRUE)
```

Subset data and remove any taxa not in subset:

```
filter_obs(x, "info", n_legs == 4, drop_taxa = TRUE)
```

Add a column to a dataset:

```
mutate_obs(x, "info", bipedal = n_legs == 2)
```

pytaxa: the Python package

github.com/sckott/pytaxa docs

`pip install pytaxa`

in development: porting `taxa` R client to Python

```
from pytaxa import examples
ex = examples.eg_hierarchy("salmo")

<Hierarchy>
  Salmonidae / family / 161931
  Salmo / genus / 161994
  Salmo salar / species / 161996
  Chordata / phylum / 158852
  Vertebrata / subphylum / 331030
  Teleostei / class / 161105
```

```
ex.pick(names = ["Salmo", "Chordata", "Teleostei"])

<Hierarchy>
  Salmo / genus / 161994
  Chordata / phylum / 158852
  Teleostei / class / 161105
```

taxa/pytaxa plans

- pytaxa parity with taxa
- shuttle data between pytaxa & taxa?
  - protocol buffers?
- arbitrary data backends? (e.g., sqlite db of taxonomy or user data)
- mappings to Darwin Core terms

https://github.com/ropensci/taxa

https://github.com/sckott/pytaxa