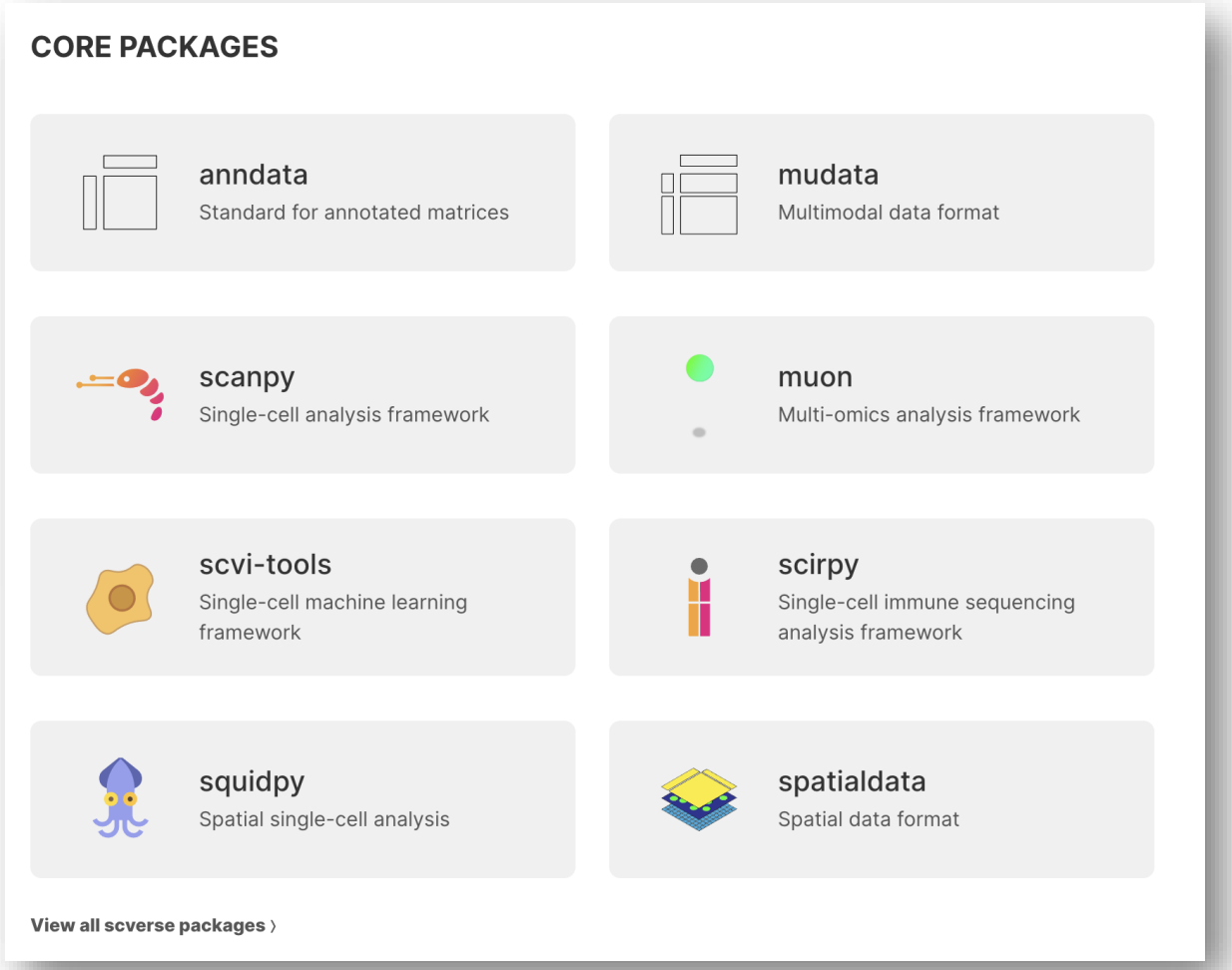
https://scverse.org

単一細胞オミクス解析に関連するPythonツールの開発・維持を目的に2022年に組織されたコンソーシアム。

AnnDataとScanpyをコア技術とする。
本日扱うscVelo, CellRankもコア技術を土台に開発された
Ecosystem packageのうちのひとつ。

マルチモーダルデータ（scRNA-seq + scATAC-seq）の解析に対する拡張として MuData, Muon の開発、
空間トランスクリプトーム解析のためのSquidpyの開発など。
それぞれの相互運用性の改善やファイルフォーマットの統一など、
一体として扱いやすいツール群の開発を目指していくコミュニティ。



AnnData

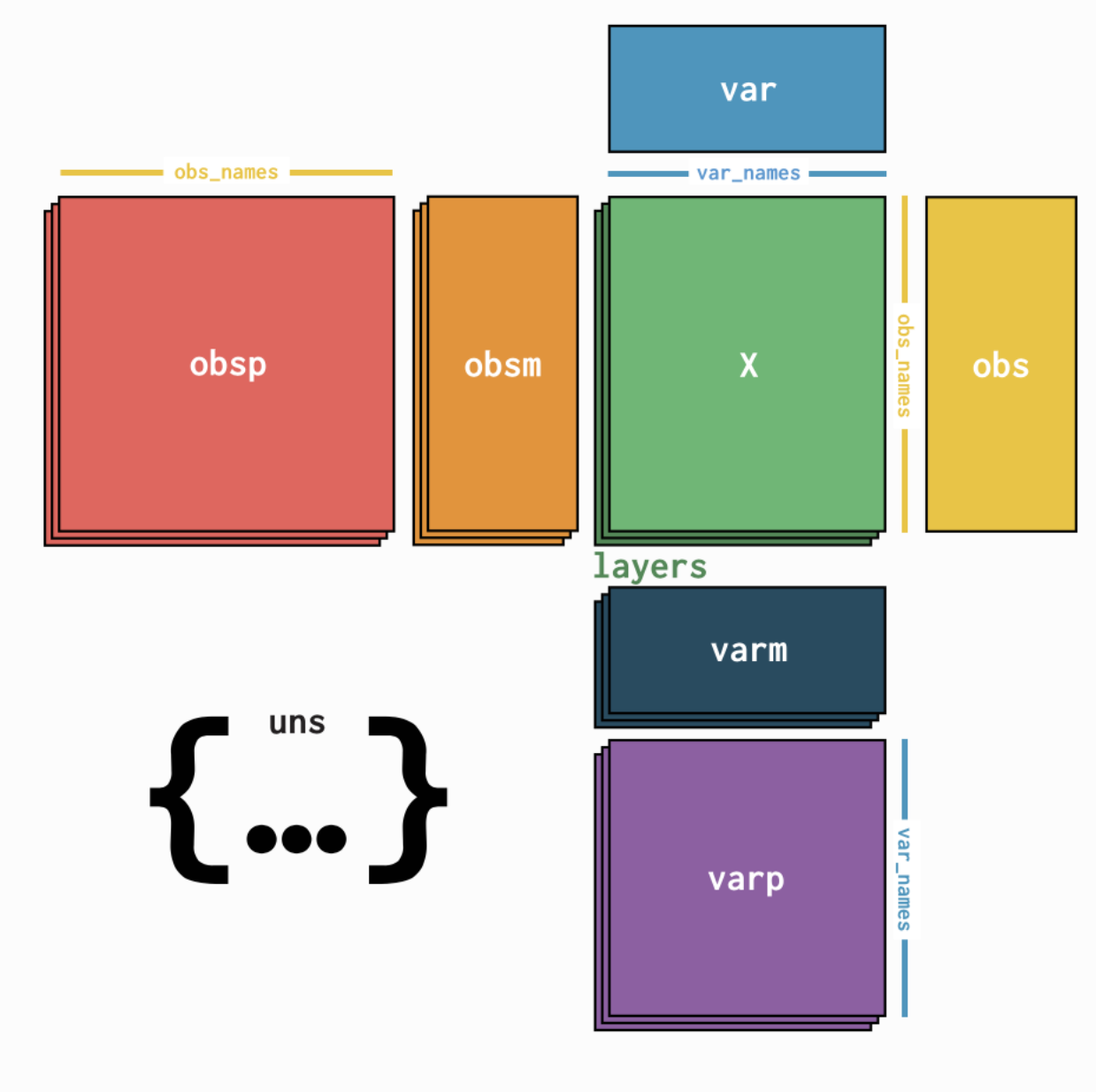
“Annotated Data”（アノテーションされたデータ）の略。

オミクスデータ格納のためにPandasのDataFrameを拡張したデータ構造。
シングルセル解析のためのPythonパッケージの多くが、このオブジェクトに対する計算として実装されている。

オミクスデータは実験で測定された数値テーブルのほかに、
観測値（obs）、変数（var）それぞれが多様な情報を持つ。
たとえばRNA-seqの場合、観測値であるサンプルは実験条件・性別・年齢など
様々なメタデータを持つ。変数である遺伝子も、遺伝子IDやシンボルだけでなく、
機能カテゴリや、DEGか否かなどのメタデータを持つ。

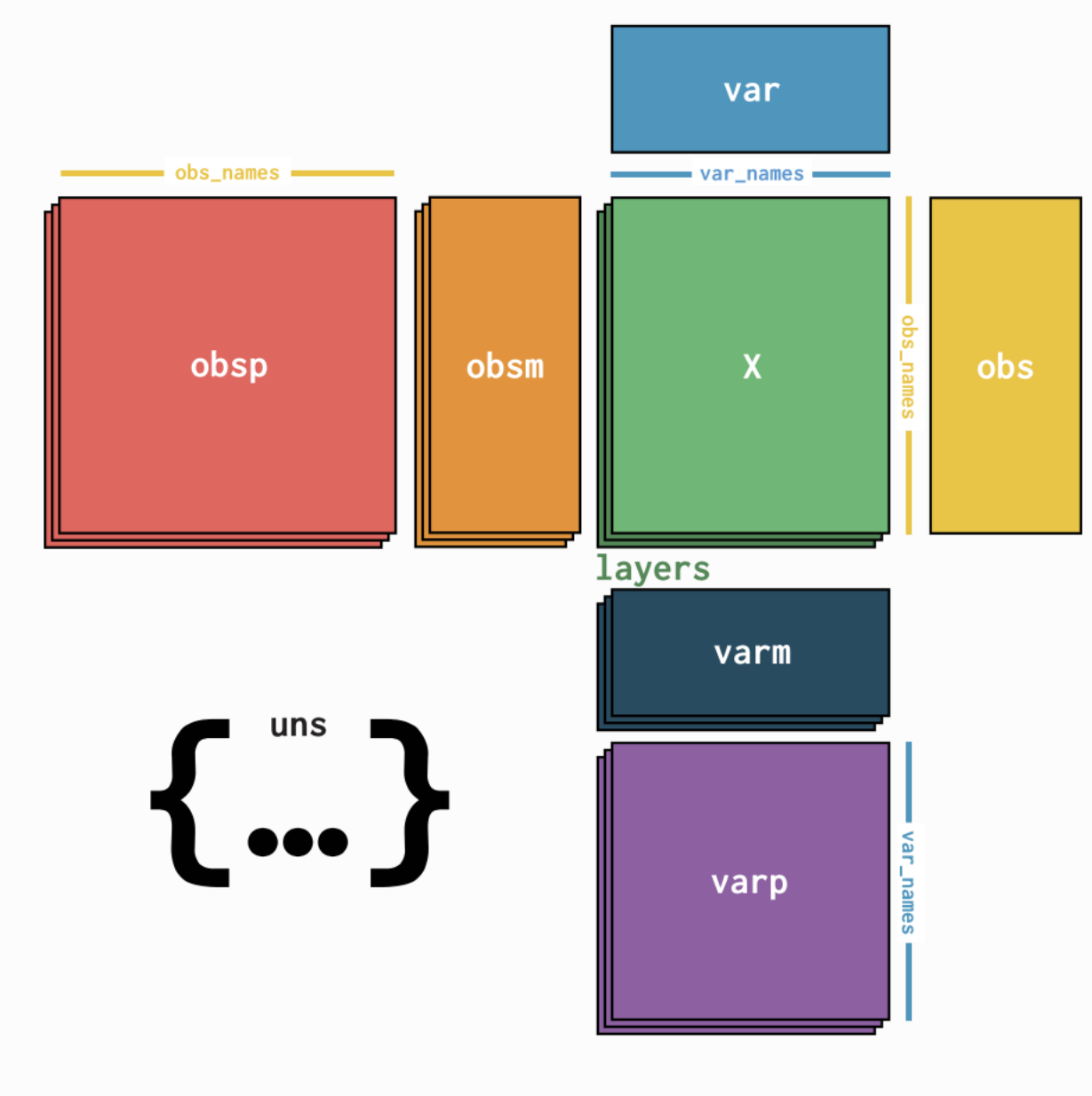
それぞれを個別のオブジェクトとして管理するのはとても面倒。
数値テーブルになんらかの操作を施した結果が、観測値や変数のメタデータに即座に反映されない。
なので、複数のオブジェクトをいちいち行ったり来たりしなきゃならない。
テーブルに対する計算の結果わかったことを観測値のメタデータに入れて、
その結果に基づいて観測値をセレクションしたから今度は数値テーブルを同じようにスライスして、、、みたいな。

そういった面倒を避けるために、すべての観測と計算結果をひとつのオブジェクトに詰め込んで管理しやすくしたのが、AnnData というオブジェクトの特徴。



AnnData

- **.X**
  n_obs × n_varsの数値テーブル。numpy.ndarrayやscipyのスパースマトリックス。scRNA-seqのカウントマトリックスなど、実験の根幹となるデータ。**layers** に、同じshapeの複数のマトリックスを保持しておける。たとえば全体をノーマライズしたけど元々のカウントデータも残しておきたいときは別のレイヤーに入れておく。スライスの影響はすべてのlayerに作用する。
- **.obs**
  observationsの略。観測値に関するメタデータ。PandasのDataFrameなのでPandasの操作は全部実行できる。長さは必ず n_obs
- **.var**
  variablesの略。変数（遺伝子など）に関するメタデータ。PandasのDataFrame。長さは必ず n_var
- **.obsm**
  multi-dimensional annotations for obs. 複数の数値のまとまりでそれぞれの観測値を表現したいときに使う。各観測値の低次元空間座標など。次元サイズは任意。n_obs × 次元サイズの numpy.ndarray.
- **.varm**
  multi-dimensional annotations for var. n_var × 次元サイズのnumpy.ndarray
- **.obsp**
  Pairwise annotation of obs. 観測値のペアに関する情報。距離行列など。n_obs × n_obs のnumpy.ndarray
- **.varp**
  Pairwise annotation of var. 変数のペアに関する情報。距離行列など。n_var × n_var のnumpy.ndarray
- **.uns**
  それ以外のデータ。とくに構造の制限はない。その他の関連データをひとまとめにしておきたいときに辞書型で放り込んでおく。クラスタの色指定とか。



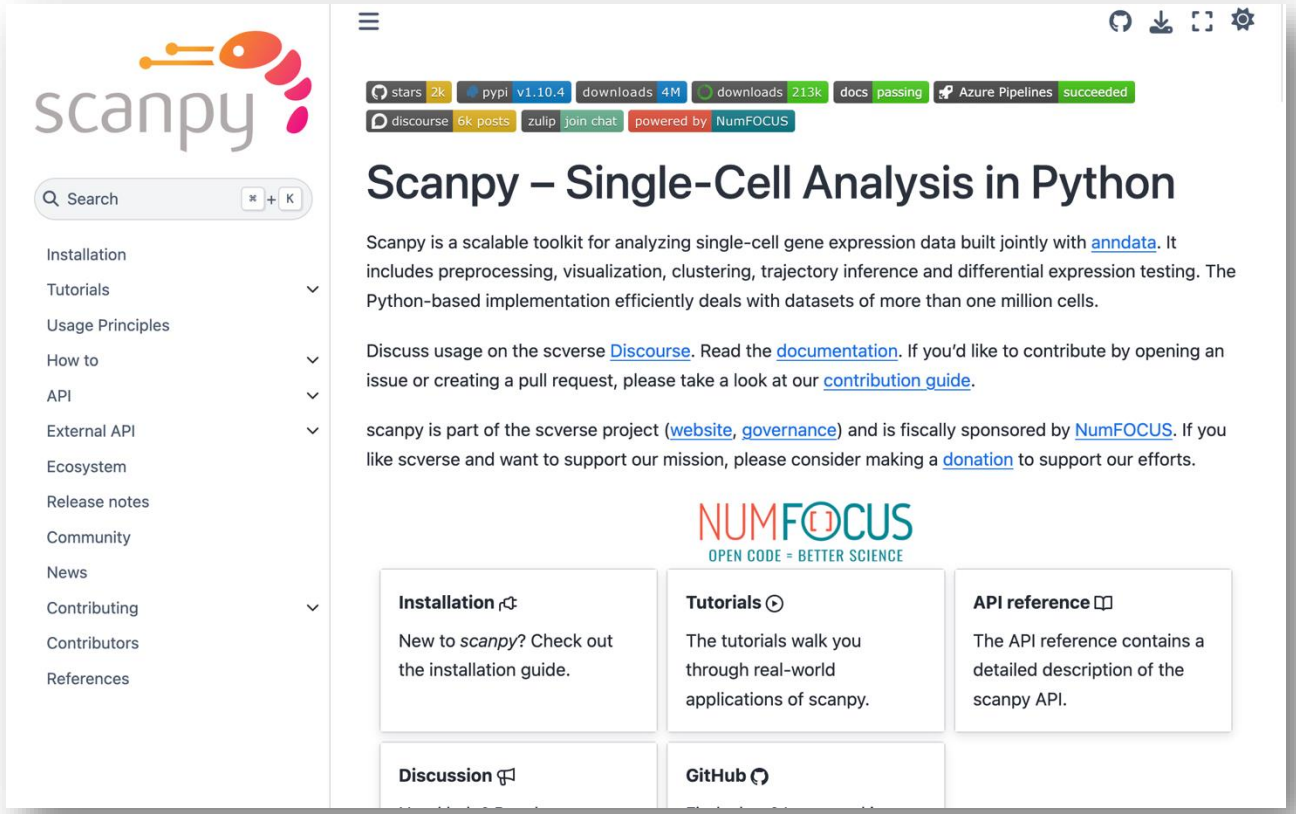
Scanpy

Pythonでシングルセル解析をする際のコアパッケージ。

データの前処理や、近傍グラフ構築、t-SNEなど、
標準的な解析を実行できる。

基本的に、
AnnDataオブジェクトを入力して関数を実行すると、
結果が同じAnnDataオブジェクトに追加されていく。
新しいAnnDataを返すのではなく、
inplaceで（＝破壊的に）AnnDataが変換されていくのが特徴。

一見どこにどんな変化が生じたのかわかりにくい。
観測値や変数のデータフレームにいつのまにか勝手にカラムが追加されていることがある。



https://scanpy.readthedocs.io

Scanpyの関数

- **scanpy.pp.XXX**
  前処理（**p**re**p**rocessing）に関連する関数がある。
  細胞や遺伝子のフィルタリング、対数変換や、近傍グラフの構築など

- **scanpy.tl.XXX**
  さまざまなツール（**t**oo**l**s）のセット。
  PCA, t-SNE, UMAPなどの次元削減や、Louvain/Leidenクラスタリングなど。

- **scanpy.pl.XXX**
  プロット（**pl**otting）用の関数。
  PCA用のプロット、UMAP用のプロットなど、それぞれの可視化に適した関数が用意されている。
  複雑な処理を書かなくても、anndataに含まれるメタデータから自動的に、遺伝子発現量による色のグラデーションや、クラスタごとの色分けなどをやってくれる。

注：scanpyはたいてい “sc” の短縮名で呼び出すことが多いので、以上の関数は、sc.pp.XXX, sc.tl.XXXなどと呼び出す