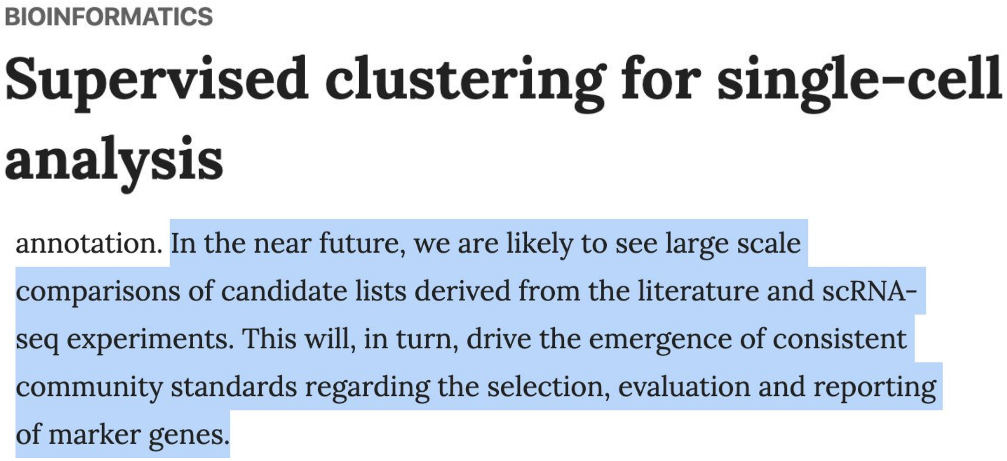
CHEM 545/546 Technology Review

Nicolas Cardozo
Steven Wu
Samantha Borje

Background

- The problem we are trying to solve is curating a list of single cell RNA-sequencing markers from across all cell types.

- We are going curate the corpus from millions of abstracts from pubmeds API.



BIOINFORMATICS

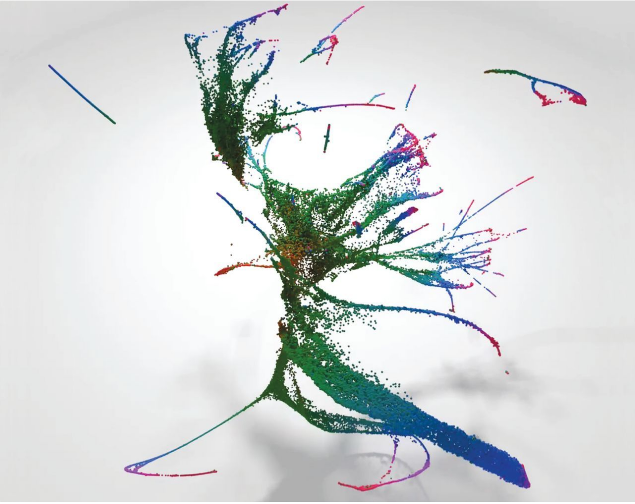
Supervised clustering for single-cell analysis

annotation. In the near future, we are likely to see large scale comparisons of candidate lists derived from the literature and scRNA-seq experiments. This will, in turn, drive the emergence of consistent community standards regarding the selection, evaluation and reporting of marker genes.

Use case

- Cancer biologist and any scientists that utilize scRNA-seq experiments will use this technology.
- Using these celltype-gene word associations will integrate biological domain knowledge and data-driven discovery to identify new cell-types.

Example of plot we would like to create:



Potential python libraries

- The potential python libraries that we are thinking about using to address our technology requirements are:
    1. Natural Language Toolkit (NLTK)
    2. Gensim
    3. SpaCy
- These three libraries have been chosen because they provide:
    - Well-documented, Interactive user experience
    - Simple to use and implement in Python

Others:

NLTK

PROS

Most developed/well-known library for NLP

Third-party extensions

Different approaches for each case

CONS

Complicated to learn, use; slow

Doesn’t provide neural network models

No integrated word vectors

spaCy

PROS

Super-fast (fastest NLP library)

Highly optimised tools for each task

Active support and development

CONS

Lacks flexibility for special cases

Slow sentence tokenization

Object-oriented

Our choice:

gensim

REASONS

Best for working with large datasets; processes datastreams

Supports deep-learning; designed primarily for unsupervised text-modelling

Active support and development

CONSIDERATIONS

Doesn’t have the widest range of tools; might need to use some from Spacy or NLTK

Thanks for your attention!

Questions?