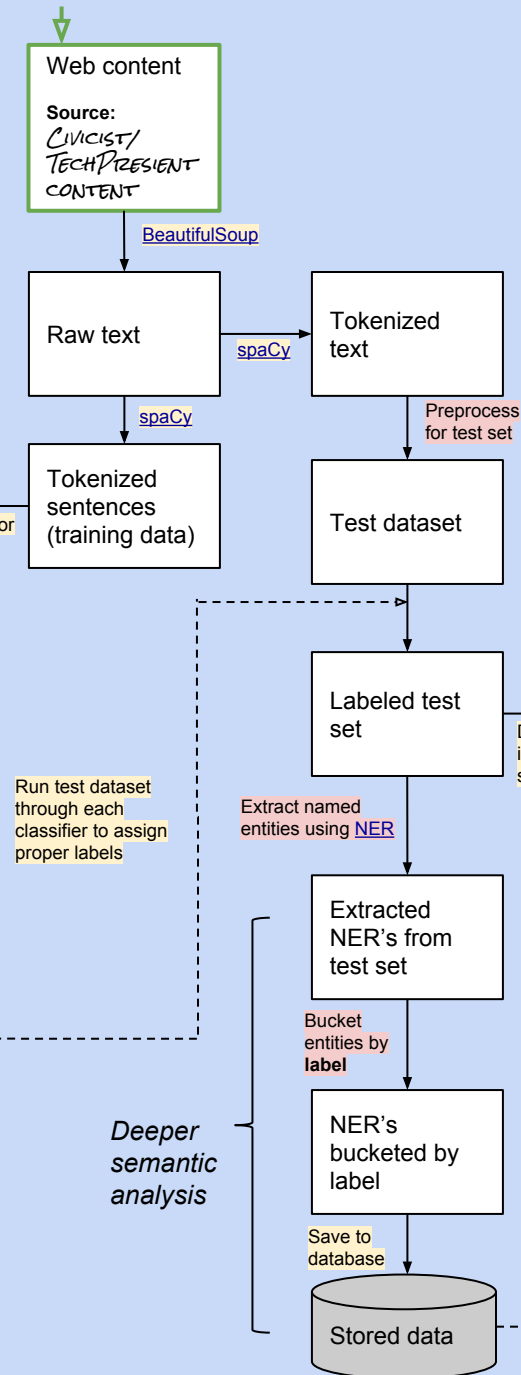# Classifiers

Labels:
(one for each connection type in Civic Graph)
- Funding
- Data
- Employment
- Collaboration
- Location
- Other/Undefined *
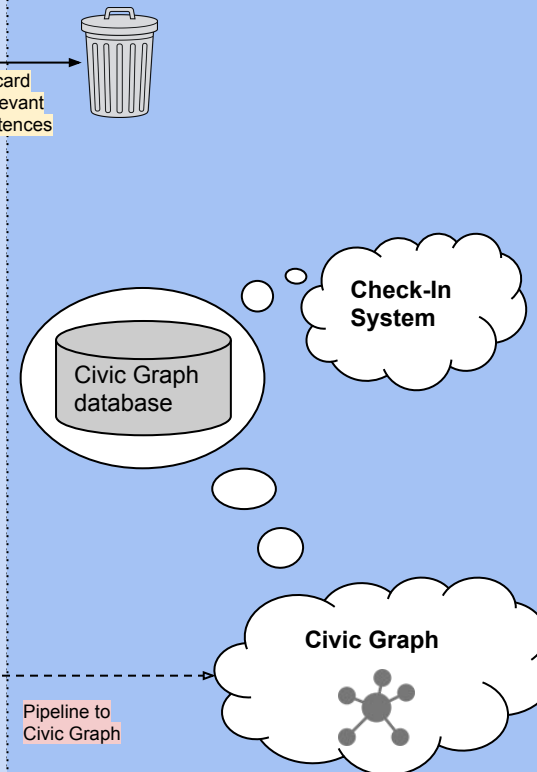
Scrapers/Tokenizers

Web content

Source:
Civicist/
TechPresent
content

BeautifulSoup

Raw text

spaCy → Tokenized text

spaCy → Tokenized sentences (training data)

Generate labels for each sentence manually (use Amazon MTurk?)

Labeled, tokenized training data

Combine labels with associated sentence

Training datasets (one per **label**)

Train classifiers using Support Vector Machine (svm)

Classifiers (one per **label**)

Run test dataset through each classifier to assign proper labels

Preprocess for test set

Test dataset

Labeled test set

Discard irrelevant sentences

Extract named entities using NER

Extracted NER's from test set

Bucket entities by **label**

Deeper semantic analysis

NER's bucketed by label

Save to database

Stored data

Pipeline to Civic Graph

Key:
- Code written
- Not done
- Source subject to change

Civic Graph database

Check-In System

Civic Graph

Pipeline to Civic Graph

Civic Graph
MAD LIBS

Automating Data Collection for Microsoft Civic Graph

(Draft Pipeline)