Biomedical text mining to discover how BCL2 genes interact with family genes to regulate apoptosis

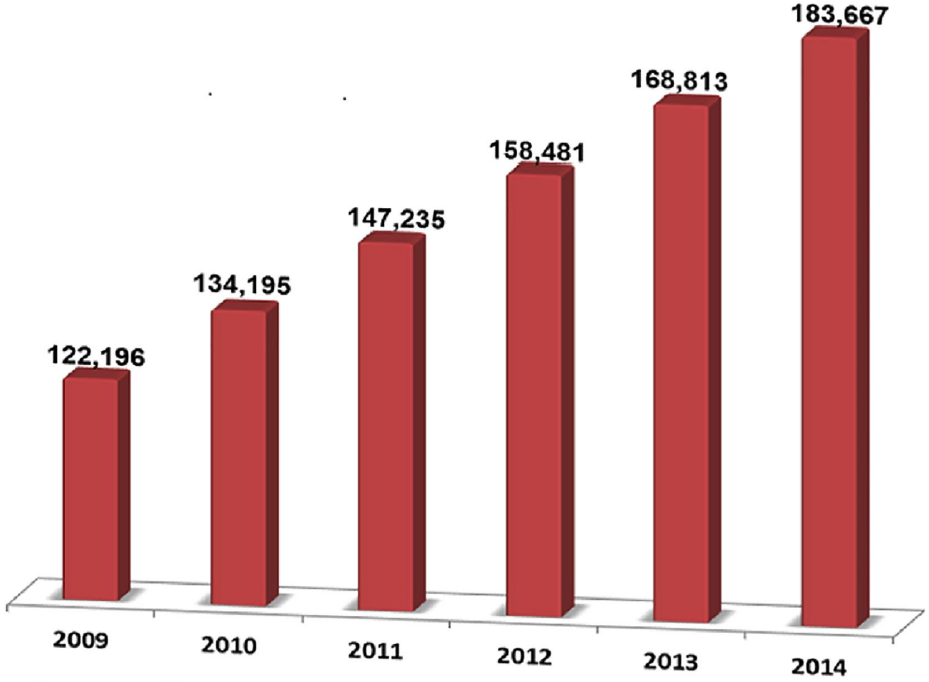
Yu Zhang & Ming Chen

COSC526

Outline

- Background
- Approaches
- Results
- Challenges
- Further analysis

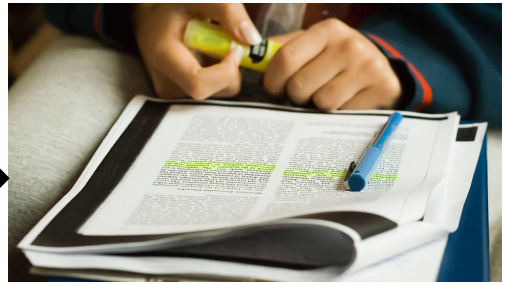
Unbelievable data harvest grows in biomedical research!



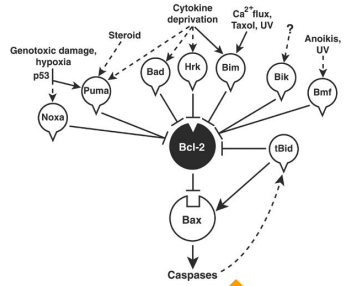
The number of publications obtained by searching the Pubmed database with keyword “cancer” (Ye et al. 2016, *plos one*).

Scientific Review 1.0 & 2.0

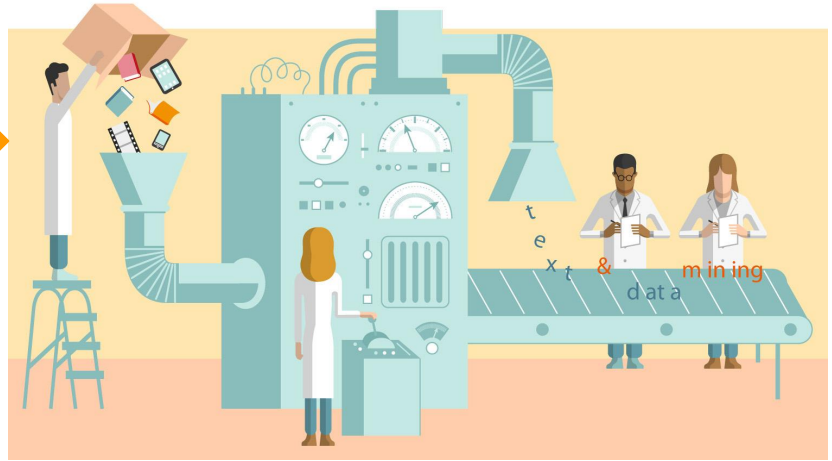
1.0



1.0



2.0



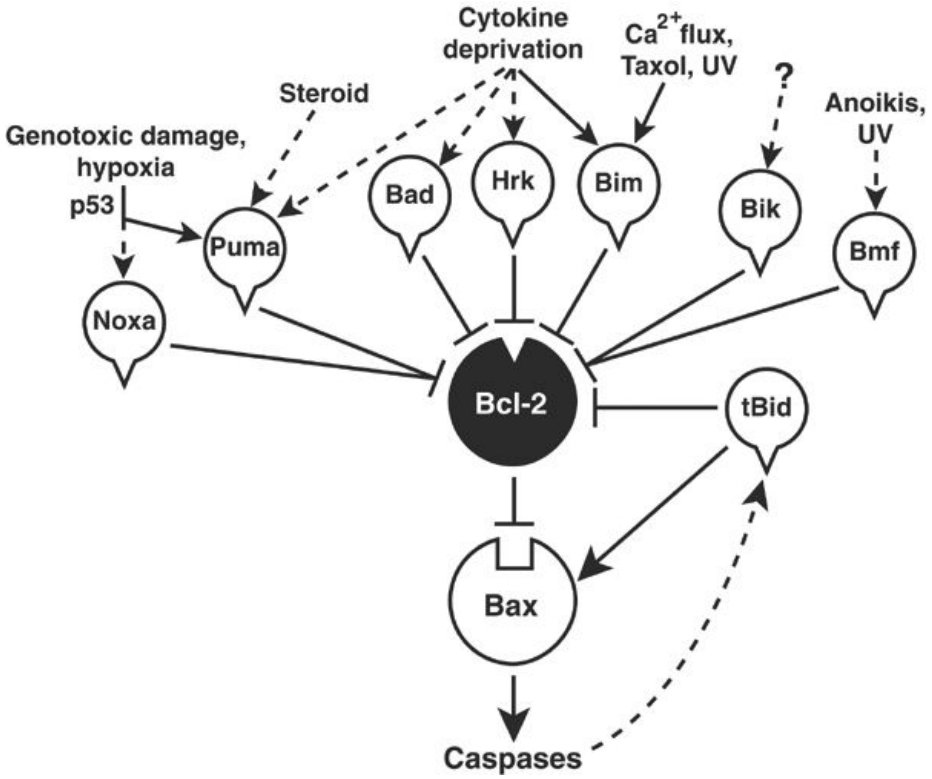
2.0

Goal:

Manage the machine to search and screen the published literature and produce a simple review for our interest.

Trail test:
Use the PubMed literature on BCL2 gene as the resource and build a simple relationship network as the review for BCL2 gene and its brother genes within the BCL2 family.

B-Cell Lymphoma 2 family and Apoptosis



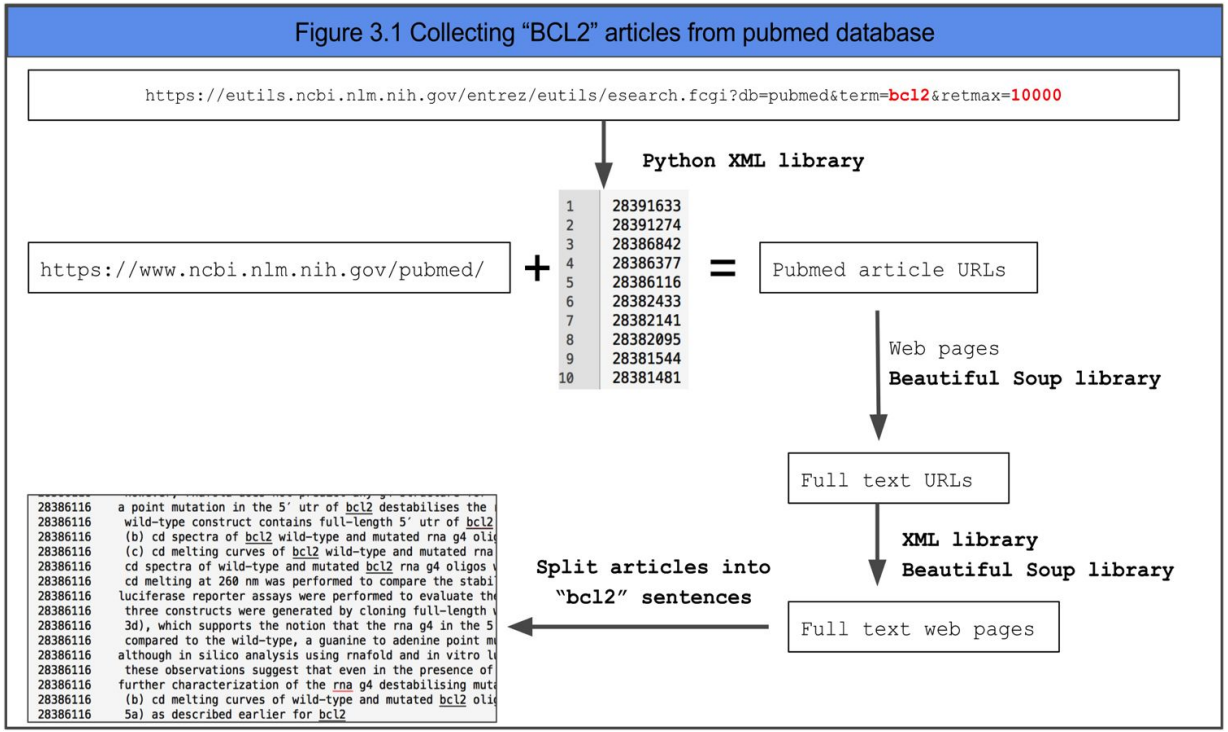
Data Collection

8109 article IDs

Only scanned 3102 article IDs

502 full text articles

5110 BCL2 sentences



Data processing: split sentences into words and lemmatize words

Figure 3.2 Convert “BCL2” sentences to lemmatized words

```
from pyspark.sql.functions import udf
from pyspark.sql.types import *
```

```
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()

def lemmatizer(s):
    words_list = [wordnet_lemmatizer.lemmatize(w, 'v') for w in s.split()]
    return(words_list)
lemmatizer_udf = udf(lemmatizer, ArrayType(StringType()))
```

```
bcl2_lemm = bcl2.select(bcl2.id, bcl2.sentences, lemmatizer_udf(bcl2.sentences)\
                        .alias('lemm_words'))
```

```
bcl2_lemm.show(5)
```

```
+--------+--------------------+--------------------+
|      id|           sentences|          lemm_words|
+--------+--------------------+--------------------+
|28386116| we experimentall...|[we, experimental...|
|28386116| to evaluate the ...|[to, evaluate, th...|
|28386116|it has been well ...|[it, have, be, we...|
|28386116| our bioinformati...|[our, bioinformat...|
|28386116| it has been prev...|[it, have, be, pr...|
+--------+--------------------+--------------------+
only showing top 5 rows
```

Collect gene Interaction terms and BCL2 family genes

MeSH database
- Keyword: BCL2

39 words

```
'downregulate',
'elevate',
'enhance',
'inactivate',
'increase',
'induce',
'inhibit',
'initiate',
'interact',
'interference',
'mediate',
'modulate',
'prevent',
'promote',
```

23 genes

```
'bad',
'bak',
'bax',
'bcl-2a1',
'bcl-b',
'bcl-w',
'bcl-xl',
'bcl-xs',
'bfl-1',
'bid',
'bik',
'bim',
```

BCL2 family genes

Filter sentences containing “BCL2”, at least one gene interaction terms and at least one family gene

```
def filter_bcl2_regulation(l):
    set0 = set(gr_lemm_words)
    set1 = set(l)
    common_words = list(set0.intersection(set1))
    if len(common_words) > 0:
        return(common_words)
    else:
        return(None)
filter_bcl2_regulation_udf = udf(filter_bcl2_regulation, ArrayType(StringType()))
```

```
bcl2_regulation_df = bcl2_lemm.select(bcl2_lemm.id,
                                      bcl2_lemm.sentences,
                                      bcl2_lemm.lemm_words,
                                      filter_bcl2_regulation_udf(bcl2_lemm.lemm_words)\
                                      .alias('bcl2_regulation'))
```

```
bcl2_regulation_df = bcl2_regulation_df.filter(bcl2_regulation_df.bcl2_regulation.isNotNull() )
```

```
bcl2_regulation_df.show(5)
```

```
+--------+--------------------+--------------------+----------------+
|      id|           sentences|          lemm_words| bcl2_regulation|
+--------+--------------------+--------------------+----------------+
|28386116|it has been well ...|[it, have, be, we...|[proto-oncogene]|
|28386116| compared to the ...|[compare, to, the...|      [increase]|
|28386116|bcl2 is a human p...|[bcl2, be, a, hum...|[proto-oncogene]|
|28386116| many examples ex...|[many, examples, ...]|       [elevate]|
|28386116| several mechanis...|[several, mechani...|[overexpression]|
+--------+--------------------+--------------------+----------------+
only showing top 5 rows
```

```
def filter_bcl2_family(l):
    set0 = set(bcl2_family_lemm)
    set1 = set(l)
    common_words = list(set0.intersection(set1))
    if len(common_words) > 0:
        return(common_words)
    else:
        return(None)
filter_bcl2_family_udf = udf(filter_bcl2_family, ArrayType(StringType()))
```

```
bcl2_family_df = bcl2_regulation_df.select(bcl2_regulation_df.id,
                                           bcl2_regulation_df.sentences,
                                           bcl2_regulation_df.lemm_words,
                                           bcl2_regulation_df.bcl2_regulation,
                                           filter_bcl2_family_udf(bcl2_regulation_df.lemm_words)
                                           .alias('bcl2_family'))
```

```
bcl2_family_df = bcl2_family_df.filter(bcl2_family_df.bcl2_family.isNotNull())
```

```
bcl2_family_df.show(5)
```

```
+--------+--------------------+--------------------+--------------------+-----------+
|      id|           sentences|          lemm_words|     bcl2_regulation|bcl2_family|
+--------+--------------------+--------------------+--------------------+-----------+
|28369145| albicans or go-p...|[albicans, or, go...|          [increase]|      [bax]|
|28369145| (f) increased ra...|[(f), increase, r...|[increase, apopto...|      [bax]|
|28367088| the addition of ...|[the, addition, o...|  [increase, reduce]|      [bax]|
|28350842| it was demonstra...|[it, be, demonstr...|[anti-apoptotic, ...|      [bax]|
|28334048| also, there was ...|[also,, there, be...|[pro-apoptotic, i...|      [bax]|
+--------+--------------------+--------------------+--------------------+-----------+
only showing top 5 rows
```

Classification models: define interaction categories

Classification models: manually label a subset of data

bad		in contrast to our observations in the ft, hu et al [20] reported that cigarette smoke extract has the o
bad		the observed increase in bcl2 could also be responsible for the decrease in bad transcription
bad		therefore, it is possible that cigarette smoking increases bcl2 expression and that this indirectly leads
bad		alternatively, changes in the relative levels on bad and bcl2 may promote an environment suited to e
bad		it is therefore possible that reduced bad and increased bcl2 expression in the fallopian tube, as a resu
bad		it is therefore possible that reduced bad and increased bcl2 expression in the fallopian tube, as a resu
bak	-1	in addition, p-ask1, ask1, p-jnk, jnk1, jnk2, bax, bak and bim expression levels were significantly highe
bak	-1	bh3 mimetics are designed to inhibit anti-apoptotic bcl2 family proteins, leading to bax and bak activ
bak		and the bh3-only proteins bim, bid, puma, noxa, bad, bik, bmf, and hrk, which share homology with c
bak	-1	the indirect activation model proposes that bax and bak are tonically activated but are restrained by
bak	-1	in this model, bh3-only proteins induced by various death signals primarily inhibit the anti-apoptotic
bak		in this model, bh3-only proteins induced by various death signals primarily inhibit the anti-apoptotic
bak		in this model, bh3-only proteins induced by various death signals primarily inhibit the anti-apoptotic
bak		instead, binding of bh3 mimetics to anti-apoptotic bcl2 family members must result in bax and/or ba
bak		in model 2 (right), bak and/or bax are constitutively activated and are displaced from anti-apoptotic l
bak		to the extent that bh3-only proteins are constitutively activated but sequestered by anti-apoptotic b
bak	-1	based on these observations, bh3 mimetics might be killing cells by displacing partially activated bak
bak		moreover, nelarabine combined with zstk-474 induced a dephosphorylation of akt and erk1/2 and in
bak		moreover, nelarabine combined with zstk-474 induced a dephosphorylation of akt and erk1/2 and in
bak	0	moreover, nelarabine combined with zstk-474 induced a dephosphorylation of akt and erk1/2 and in

Classification models: apache spark ML pipeline

Tokenizer	`tokenizer = Tokenizer(inputCol="sentences", outputCol="words")`
HashingTF	`hashingTF = HashingTF(inputCol=tokenizer.getOutputCol(), outputCol="features")`
IDF	`idf = IDF(minDocFreq=3, inputCol="features", outputCol="idf")`
Classifier	`rf = RandomForestClassifier(numTrees=100,maxDepth=10)`
pipeline	`pipeline = Pipeline(stages=[tokenizer, hashingTF, idf, rf])`

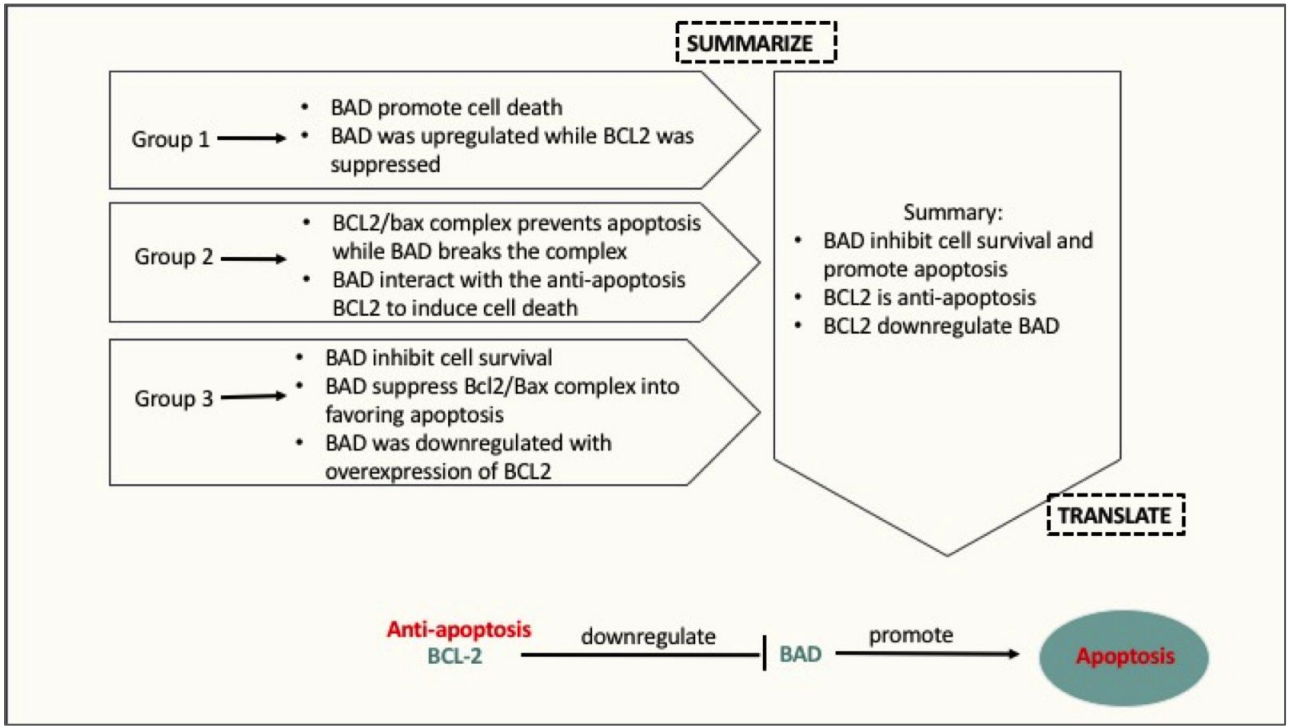
pipeline.fit(labeled data)

pipeline.transform(unlabeled data)

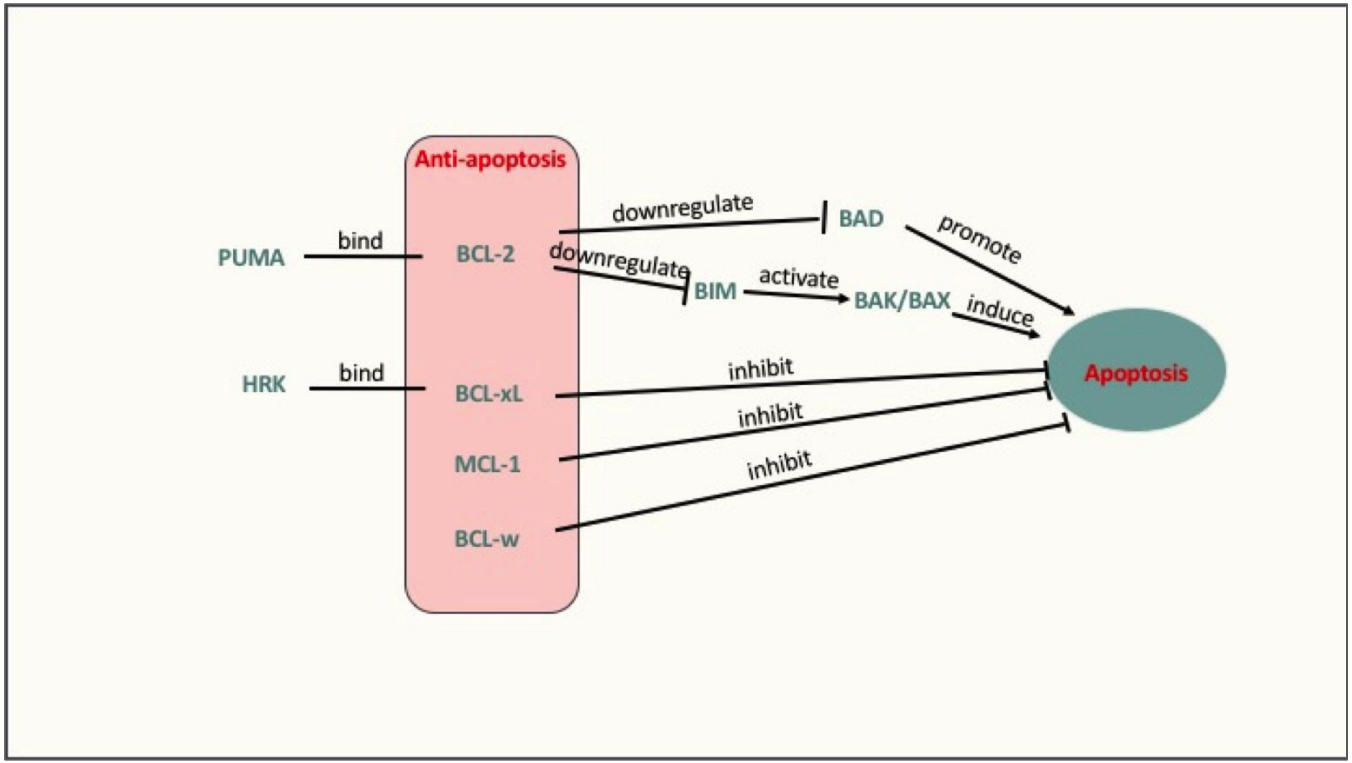
Problems

- **Very small samples for some genes**
  - e.g. 2 samples from ***a1*** gene
- **Single category**
  - Samples from most of genes only have 1 or 2 categories

Text-to-Figure translation



Regulation network centered on BCL2




APACHE
Spark


THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
BIG ORANGE. BIG IDEAS.

Questions?