Data Mining Biomedical Literature in the Cloud

(SparkText: A big data toolset for large-scale biomedical text mining)

Ahmad Pahlavan Tafti

Department of Computer Science

University of Wisconsin-Milwaukee

Mentor: Max He, Ph.D.

Marshfield Clinic®
Don't just live. **Shine.**

Outline

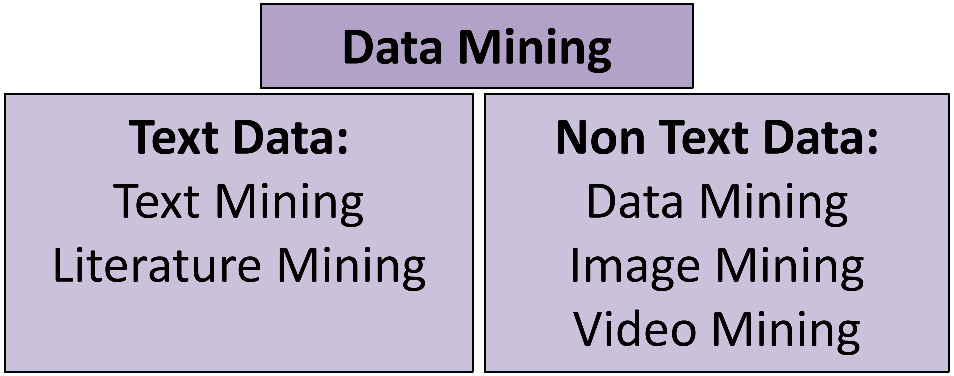
- Background
- Objective
- Methods
- Results
- Conclusions
- Future Directions

Background | Literature Mining

- **Literature mining** is used to extract information (facts or data) from text data, such as scientific articles.

- Using literature mining methods, we actually turn "**text data**" into **high-quality information** or **practical knowledge.**

- It supplies knowledge for optimal decision making.

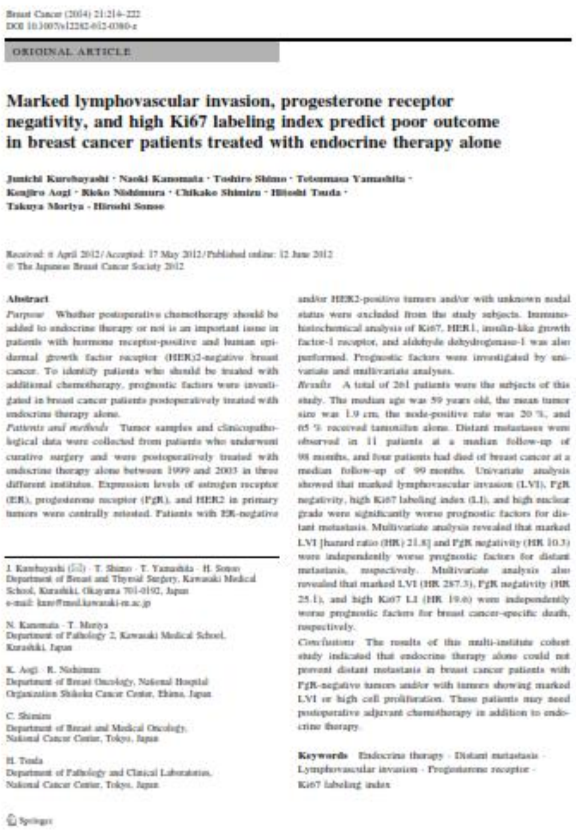


Data Mining

Text Data:
Text Mining
Literature Mining

Non Text Data:
Data Mining
Image Mining
Video Mining

Background | Literature Mining: An Application



Breast Cancer (2014) 21:214–222
DOI 10.1007/s12282-012-0380-z

ORIGINAL ARTICLE

Marked lymphovascular invasion, progesterone receptor negativity, and high Ki67 labeling index predict poor outcome in breast cancer patients treated with endocrine therapy alone

Junichi Kurebayashi · Naoki Kanomata · Toshiro Shimo · Tetsumasa Yamashita ·
Kenjiro Aogi · Rieko Nishimura · Chikako Shimizu · Hitoshi Tsuda ·
Takuya Moriya · Hiroshi Sonoo

Received: 6 April 2012 / Accepted: 17 May 2012 / Published online: 12 June 2012

Abstract

Purpose Whether postoperative chemotherapy should be added to endocrine therapy or not is an important issue in patients with hormone receptor-positive and human epidermal growth factor receptor (HER)2-negative breast cancer. To identify patients who should be treated with additional chemotherapy, prognostic factors were investigated in breast cancer patients postoperatively treated with endocrine therapy alone.

Patients and methods Tumor samples and clinicopathological data were collected from patients who underwent curative surgery and were postoperatively treated with endocrine therapy alone between 1999 and 2003 in three different institutes. Expression levels of estrogen receptor (ER), progesterone receptor (PgR), and HER2 in primary tumors were centrally retested. Patients with ER-negative and/or HER2-positive tumors and/or with unknown nodal status were excluded from the study subjects. Immunohistochemical analysis of Ki67, HER1, insulin-like growth factor-1 receptor, and aldehyde dehydrogenase-1 was also performed. Prognostic factors were investigated by univariate and multivariate analyses.

Results A total of 261 patients were the subjects of this study. The median age was 59 years old, the mean tumor size was 1.9 cm, the node-positive rate was 20 %, and 65 % received tamoxifen alone. Distant metastases were observed in 11 patients at a median follow-up of 98 months, and four patients had died of breast cancer at a median follow-up of 99 months. Univariate analysis showed that marked lymphovascular invasion (LVI), PgR negativity, high Ki67 labeling index (LI), and high nuclear grade were significantly worse prognostic factors for distant metastasis. Multivariate analysis revealed that marked LVI [hazard ratio (HR) 21.8] and PgR negativity (HR 10.3) were independently worse prognostic factors for distant metastasis, respectively. Multivariate analysis also revealed that marked LVI (HR 287.3), PgR negativity (HR 25.1), and high Ki67 LI (HR 19.6) were independently worse prognostic factors for breast cancer-specific death, respectively.

Conclusions The results of this multi-institute cohort study indicated that endocrine therapy alone could not prevent distant metastasis in breast cancer patients with PgR-negative tumors and/or with tumors showing marked LVI or high cell proliferation. These patients may need postoperative adjuvant chemotherapy in addition to endocrine therapy.

Keywords Endocrine therapy · Distant metastasis · Lymphovascular invasion · Progesterone receptor · Ki67 labeling index

J. Kurebayashi (✉) · T. Shimo · T. Yamashita · H. Sonoo
Department of Breast and Thyroid Surgery, Kawasaki Medical School, Kurashiki, Okayama 701-0192, Japan
e-mail: kure@med.kawasaki-m.ac.jp

N. Kanomata · T. Moriya
Department of Pathology 2, Kawasaki Medical School, Kurashiki, Japan

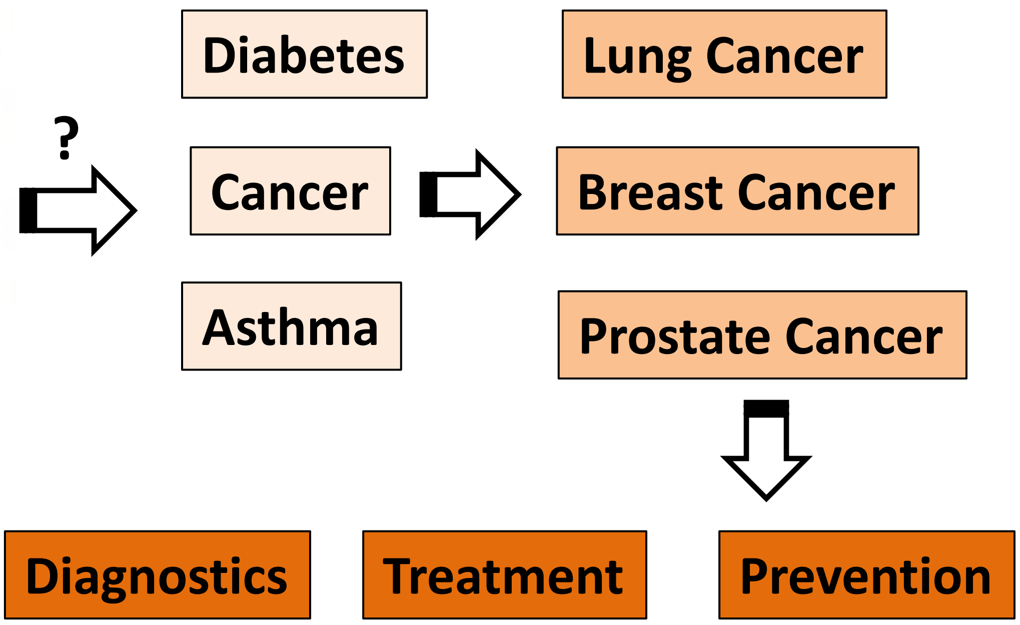
K. Aogi · R. Nishimura
Department of Breast Oncology, National Hospital Organization Shikoku Cancer Center, Ehime, Japan

C. Shimizu
Department of Breast and Medical Oncology, National Cancer Center, Tokyo, Japan

H. Tsuda
Department of Pathology and Clinical Laboratories, National Cancer Center, Tokyo, Japan

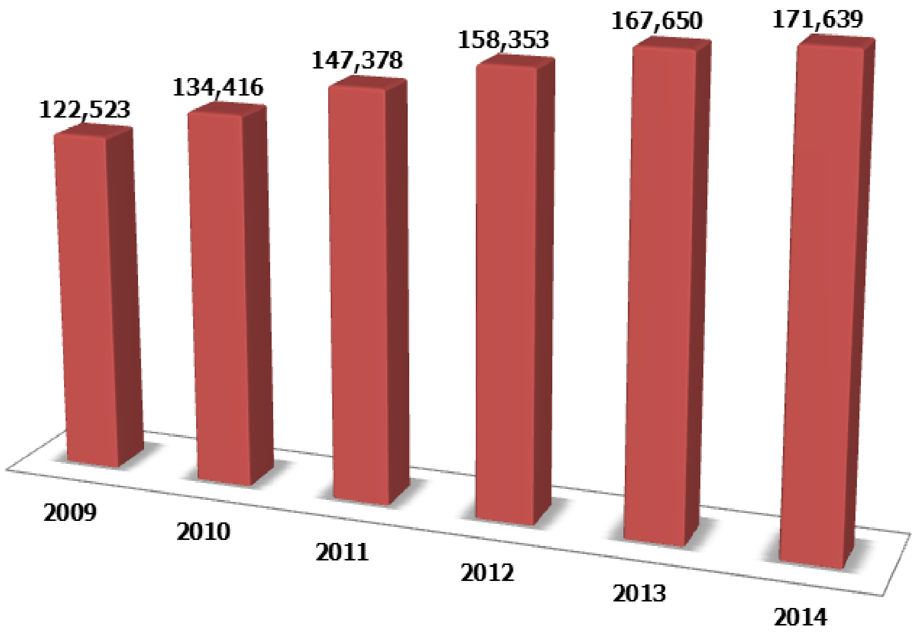
Springer

What is the subject of this paper?



Background| *Big Data in* Biomedical Literature

- We have ***Big Data*** in scientific publications in biomedical research.

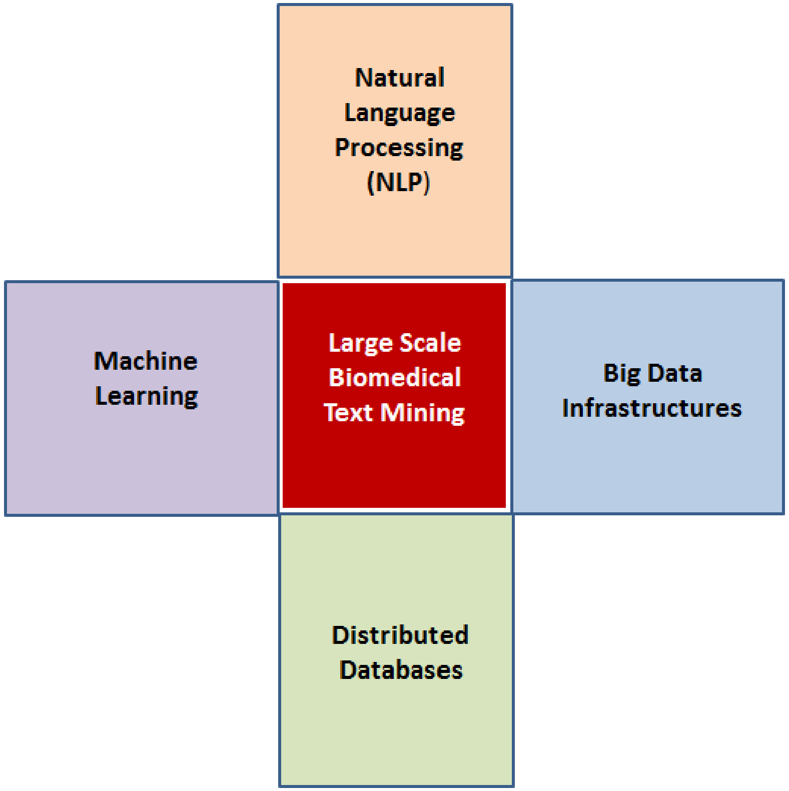


Big Data: A data set that exceed the boundaries and sizes of normal processing capabilities.

The number of publications on **Cancer research** within the last six years. Results obtained by submitting the query "cancer" in the title or abstract from **PubMed Central** (http://www.ncbi.nlm.nih.gov/pmc/). We searched "cancer" in [Abstract/Full Text/Free Full Text].

Background | Large Scale Biomedical Text Mining

- Natural Language Processing (NLP)
- Machine learning
- *Big Data* infrastructures
- Distributed Database systems



Natural Language Processing (NLP)

Machine Learning

Large Scale Biomedical Text Mining

Big Data Infrastructures

Distributed Databases

Marshfield Clinic
Don't just live. Shine.

Objective

- To develop **a scalable text mining framework for cancer research** to first extract information (e.g., breast, prostate, or lung cancers), and then develop prediction models to classify information extracted from tens of thousands of published biomedical articles downloaded from **PubMed Central** (http://www.ncbi.nlm.nih.gov/pmc/).

- **Account for:**
  - Big data set
  - ***Big Data*** infrastructures
  - Scalable machine learning techniques

Open Source Text Classification Tools

- **Compare to:**
  - **Weka Library**

    A very well-known data mining library developed in Java. (http://www.cs.waikato.ac.nz/ml/weka/)

  - **TagHelper Tools**

    An open source software which supports analysis and classifications of text data.
    (http://www.cs.cmu.edu/~cprose/TagHelper.html)

Methods| Step (1): Text Preprocessing

Prognostic significance of NANOG and KLF4 for breast cancer

Marked lymphovascular invasion, progesterone receptor negativity, and high Ki67 labeling index predict poor outcome in breast cancer patients treated with endocrine therapy alone

Mutational Processes Molding the Genomes of 21 Breast Cancers

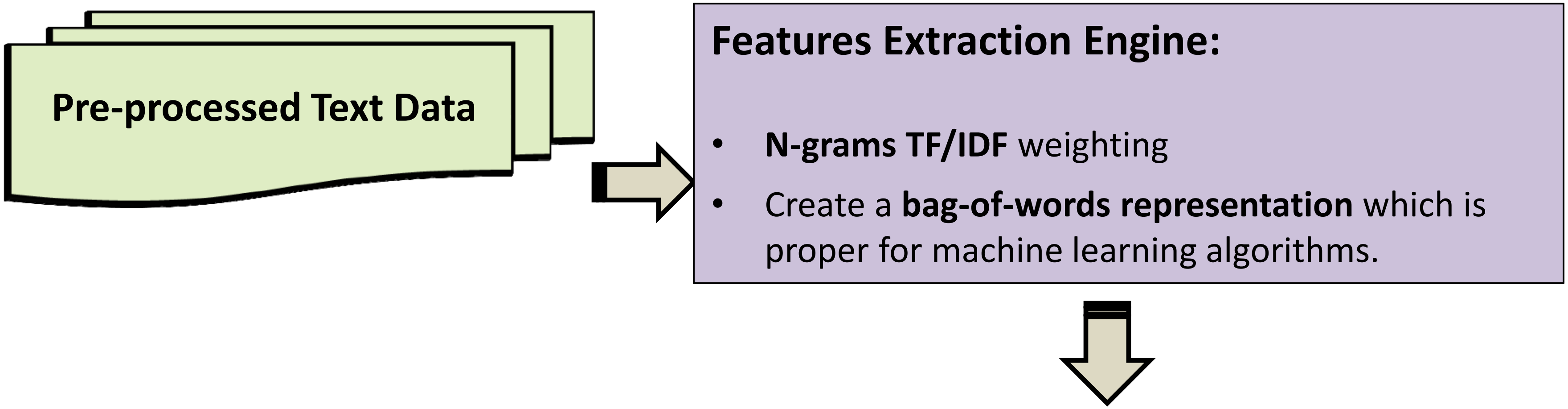
Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials

Text Preprocessing Engine:

- Replace special characters with blank spaces
- Case normalization (e.g., convert to lower case)
- Remove duplicate characters
- Remove pre-defined stop-words (e.g., “a”, “an”)
- Remove rare words
- Word stemming (e.g., “processing” → “process”)

Pre-processed Text Data

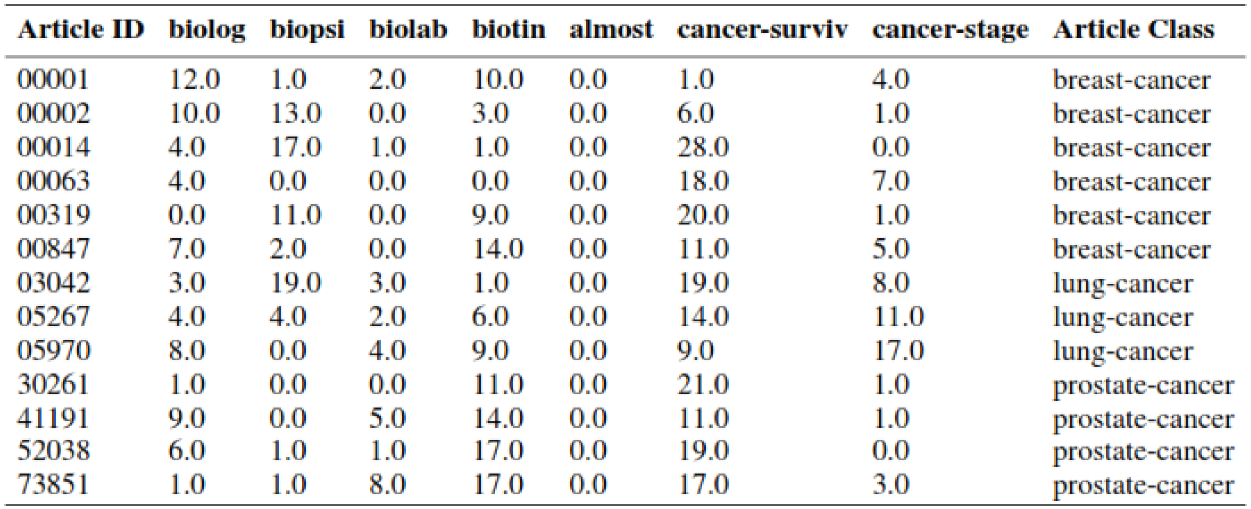
Methods| Step (2): Features Extraction (Contd.)



Pre-processed Text Data

Features Extraction Engine:

- **N-grams TF/IDF** weighting
- Create a **bag-of-words representation** which is proper for machine learning algorithms.



Article ID	biolog	biopsi	biolab	biotin	almost	cancer-surviv	cancer-stage	Article Class
00001	12.0	1.0	2.0	10.0	0.0	1.0	4.0	breast-cancer
00002	10.0	13.0	0.0	3.0	0.0	6.0	1.0	breast-cancer
00014	4.0	17.0	1.0	1.0	0.0	28.0	0.0	breast-cancer
00063	4.0	0.0	0.0	0.0	0.0	18.0	7.0	breast-cancer
00319	0.0	11.0	0.0	9.0	0.0	20.0	1.0	breast-cancer
00847	7.0	2.0	0.0	14.0	0.0	11.0	5.0	breast-cancer
03042	3.0	19.0	3.0	1.0	0.0	19.0	8.0	lung-cancer
05267	4.0	4.0	2.0	6.0	0.0	14.0	11.0	lung-cancer
05970	8.0	0.0	4.0	9.0	0.0	9.0	17.0	lung-cancer
30261	1.0	0.0	0.0	11.0	0.0	21.0	1.0	prostate-cancer
41191	9.0	0.0	5.0	14.0	0.0	11.0	1.0	prostate-cancer
52038	6.0	1.0	1.0	17.0	0.0	19.0	0.0	prostate-cancer
73851	1.0	1.0	8.0	17.0	0.0	17.0	3.0	prostate-cancer

Methods| Step (2): Features Extraction: N-grams Representation

Sentence:

The purpose of this study was to examine the incidence of breast cancer with triple negative phenotype.

Unigrams:

"The" "purpose" "of" "this" "study" "was" "to" "examine" "the" "incidence" "of" "breast" "cancer" "with" "triple" "negative" "phenotype".

Bigrams:

"The purpose" "of this" "study was" "to examine" "the incidence" "of breast" "cancer with" "triple negative" phenotype.

"purpose of" "this study" "was to" "examine the" "incidence of" "breast cancer" "with triple" "negative phenotype".

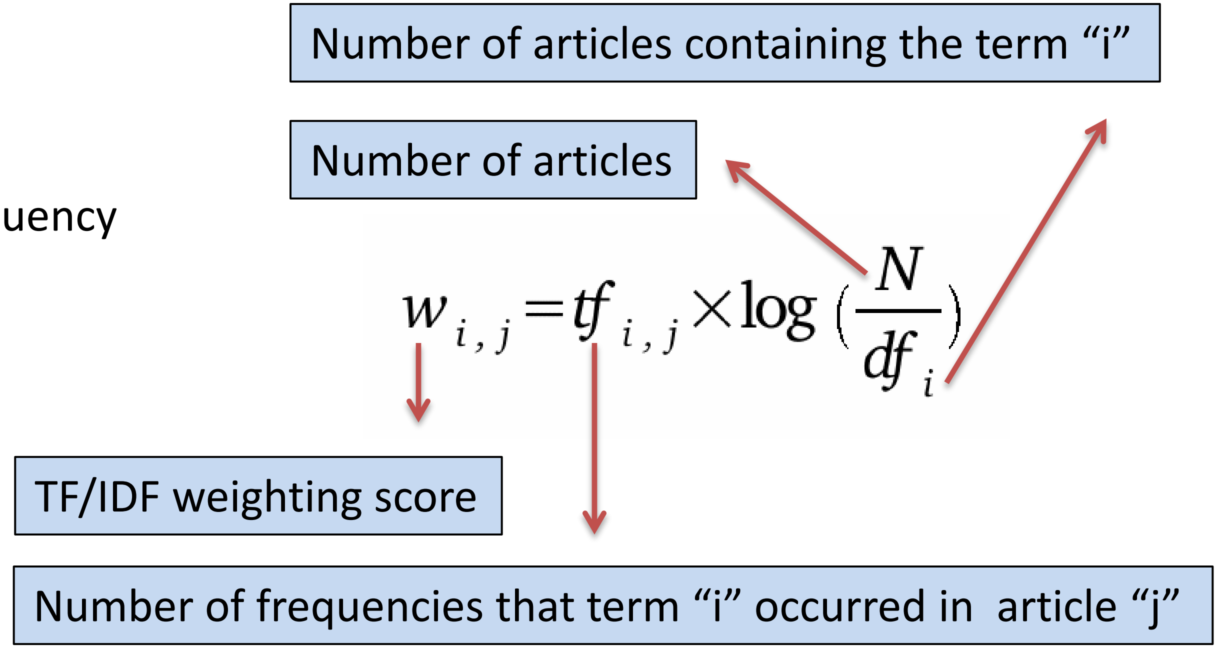
...

Marshfield Clinic
Don't just live. Shine.

Methods| Step (2): Features Extraction: TF/IDF Weighting Score

TF: Terms Frequency
IDF: Inverse Document Frequency

For a term “**i**” in article “**j**”:



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Number of articles containing the term “i”

Number of articles

TF/IDF weighting score

Number of frequencies that term “i” occurred in article “j”

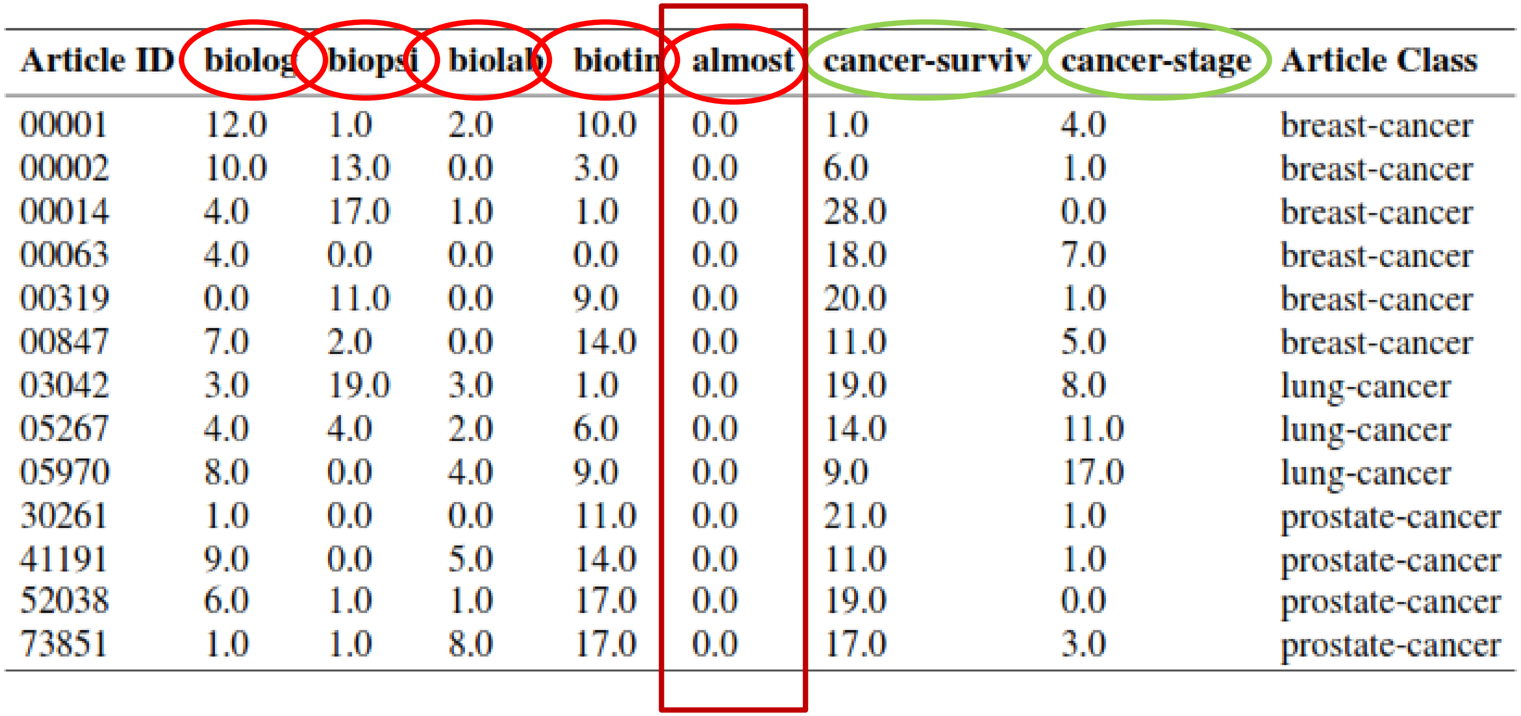
Example:
i = “almost”
j= 01 (article No. 1)
N= 100
Number of frequencies of term “almost” in article 01= 56

$W_{\text{almost, 1}}$ = 56 * log (100/100) = 56 * 0 = 0

Methods| Step (2): Features Extraction: bag-of-words Representation

unigrams

bigrams



Article ID	biolog	biopsi	biolab	biotin	almost	cancer-surviv	cancer-stage	Article Class
00001	12.0	1.0	2.0	10.0	0.0	1.0	4.0	breast-cancer
00002	10.0	13.0	0.0	3.0	0.0	6.0	1.0	breast-cancer
00014	4.0	17.0	1.0	1.0	0.0	28.0	0.0	breast-cancer
00063	4.0	0.0	0.0	0.0	0.0	18.0	7.0	breast-cancer
00319	0.0	11.0	0.0	9.0	0.0	20.0	1.0	breast-cancer
00847	7.0	2.0	0.0	14.0	0.0	11.0	5.0	breast-cancer
03042	3.0	19.0	3.0	1.0	0.0	19.0	8.0	lung-cancer
05267	4.0	4.0	2.0	6.0	0.0	14.0	11.0	lung-cancer
05970	8.0	0.0	4.0	9.0	0.0	9.0	17.0	lung-cancer
30261	1.0	0.0	0.0	11.0	0.0	21.0	1.0	prostate-cancer
41191	9.0	0.0	5.0	14.0	0.0	11.0	1.0	prostate-cancer
52038	6.0	1.0	1.0	17.0	0.0	19.0	0.0	prostate-cancer
73851	1.0	1.0	8.0	17.0	0.0	17.0	3.0	prostate-cancer

Methods| Step (3): Training and Evaluating Prediction Models

- We then apply three well-known machine learning approaches namely Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression to train and make a prediction model.

- **75%** to **train** a classifier.

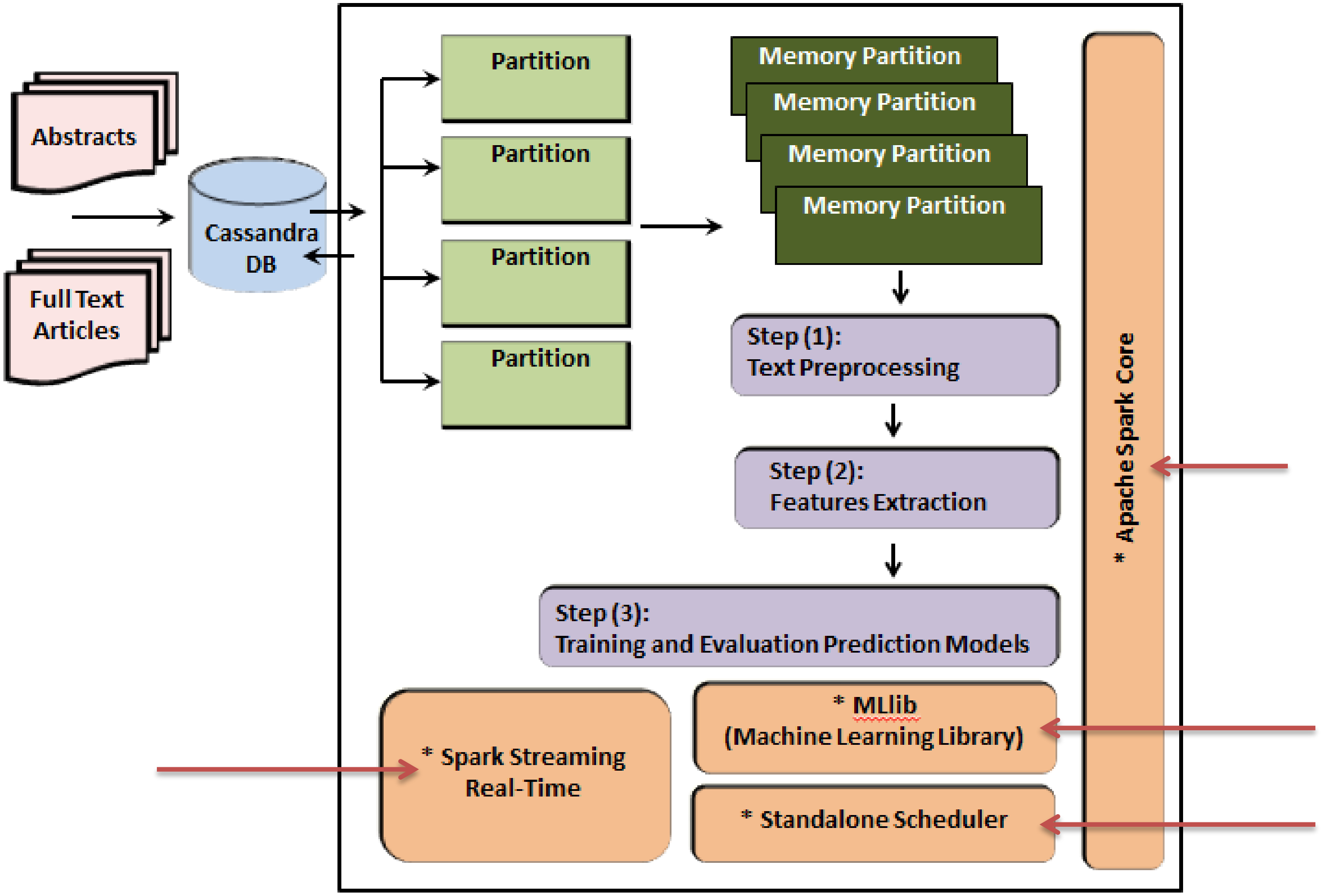
- **25%** to **test** a classifier.

Methods | Big Data Processing (Apache Spark & Cassandra DB)

- **Apache Spark™** is a fast and general engine for large-scale data processing.
  - Originally developed at AMPLab at UC Berkeley in 2009
  - Built around speed and scalability
  - Offers over 80 high-level operators to parallel apps
  - Built by a wide set of developers from over 200 companies

- **Apache Cassandra™** is an open source distributed database system designed to handle large amount of data.
  - Originally developed at Facebook.
  - Offers high availability with no single point of failure

Methods| The Block Diagram of the Proposed Framework



The framework is developed by Java programming language, Java2SE 8, and Scala.

Marshfield Clinic
Don't just live. Shine.

Results| The Dataset

Dataset	Year range	# Instances	# Breast Cancer	# Lung Cancer	# Prostate Cancer
Abstracts	**2011 – 2014**	**15983**	5476	5208	5294
Full Text I	**2011 – 2014**	**11017**	3715	3882	3420
Full Text II	**2009 - 2014**	**27001**	9118	8716	9167

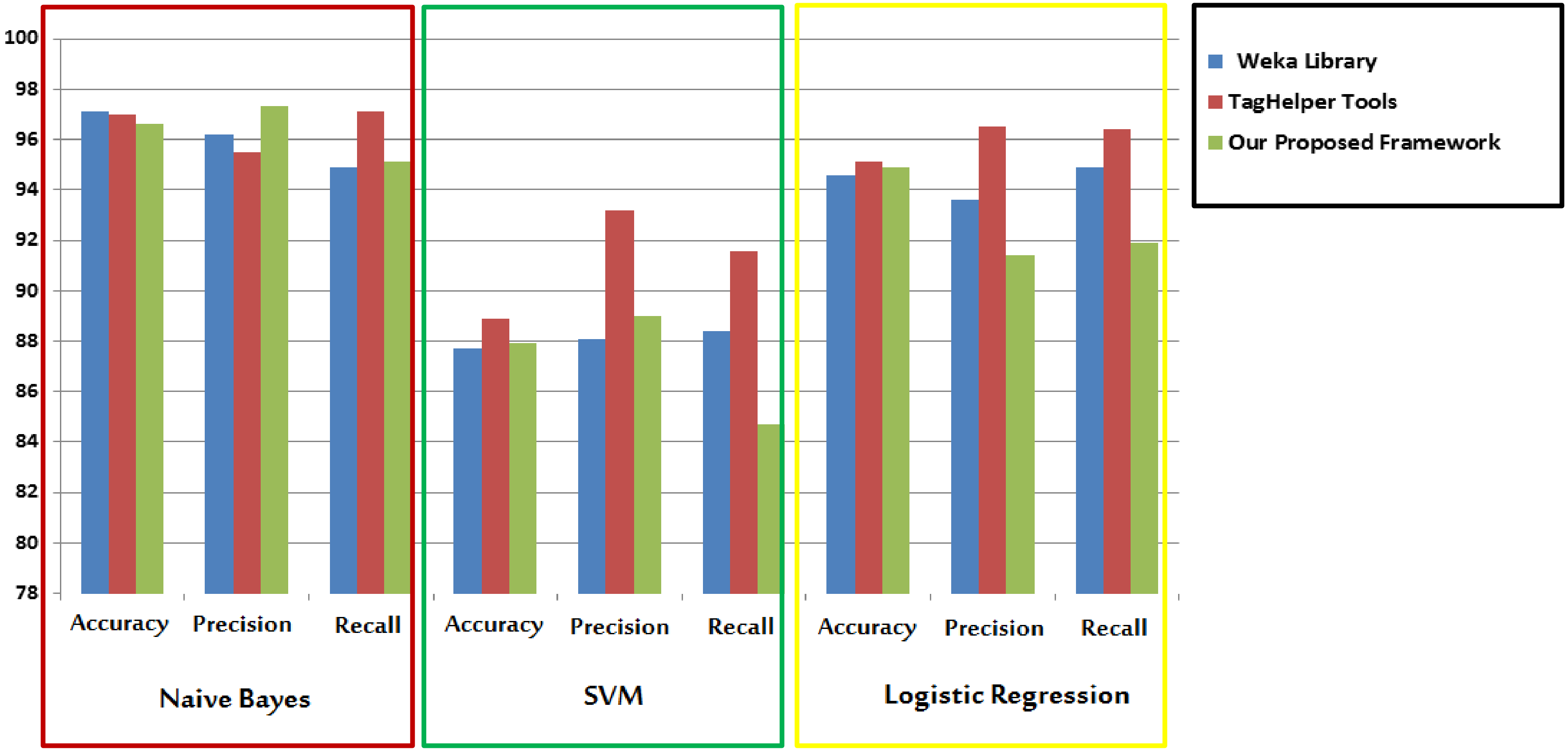
- Downloaded from **PubMed Central** (PMC) (http://www.ncbi.nlm.nih.gov/pmc/)

- For each dataset, we employed **75%** of its entire dataset to **train** the prediction model, and **25%** to **test** it.

Results| The Accuracy of the Prediction Model (Contd.)

Dataset: Abstracts
Number of instances: 15983

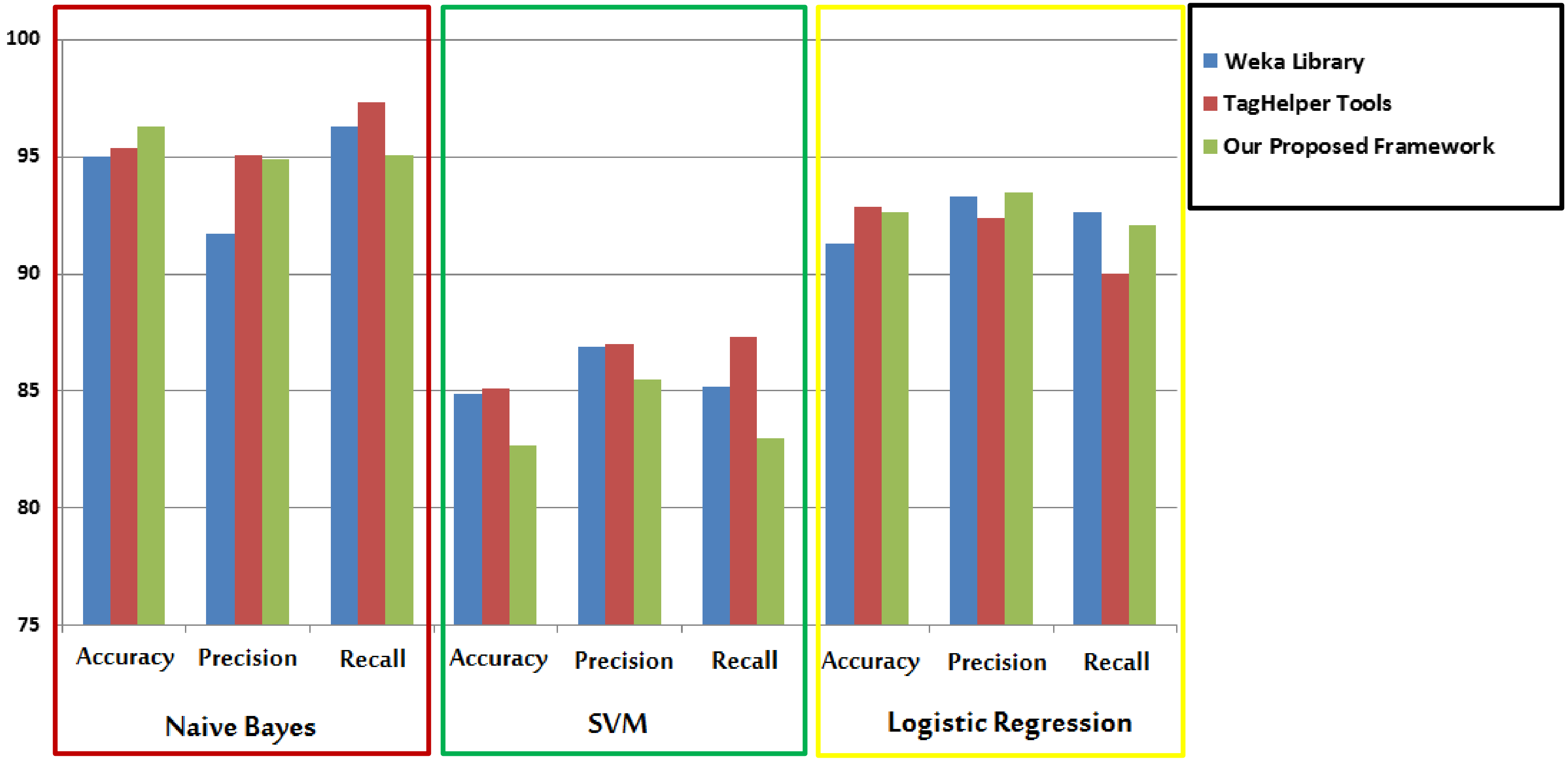
Accuracy: What percent of the prediction are correct?
Precision: What percent of positive prediction are correct?
Recall: What percent of positive cases are detected?



* Our proposed framework was developed on a *Big Data* infrastructure.

Results | The Accuracy of the Prediction Model (Contd.)

Dataset: Full Text I
Number of instances: 11017

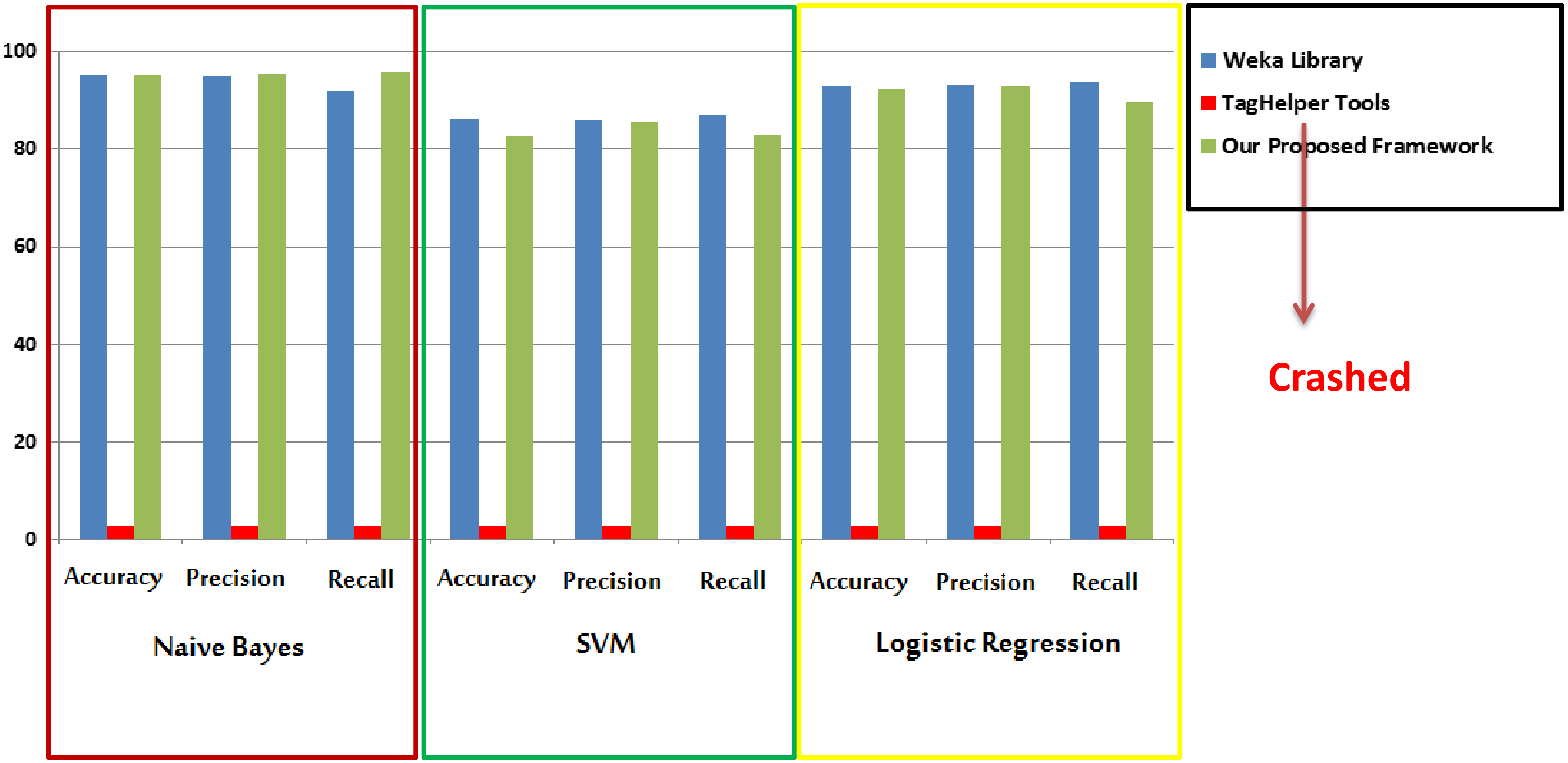


* Our proposed framework was developed on a *Big Data* infrastructure.

Marshfield Clinic
Don't just live. Shine.

Results | The Accuracy of the Prediction Model (Contd.)

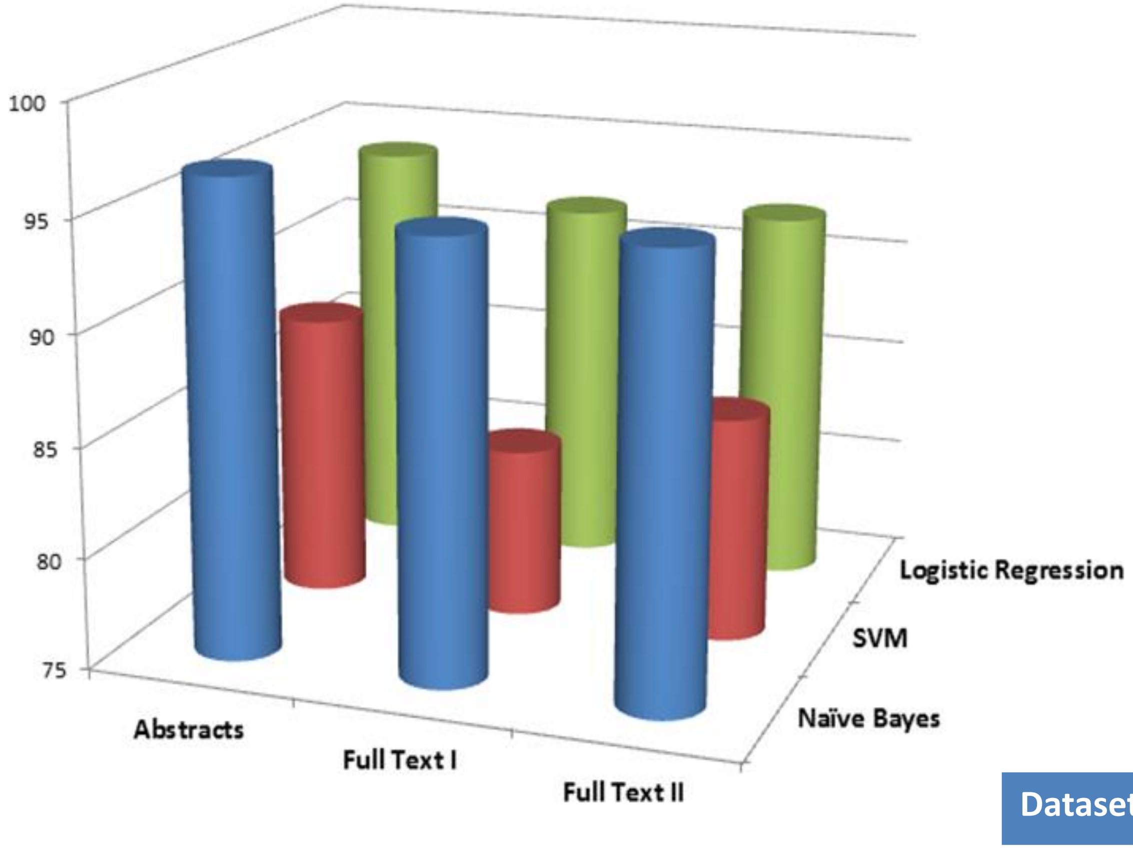
Dataset: Full Text II
Number of instances: 27001



* Our proposed framework was developed on a *Big Data* infrastructure.

Marshfield Clinic
Don't just live. Shine.

Results| The Accuracy of the proposed framework

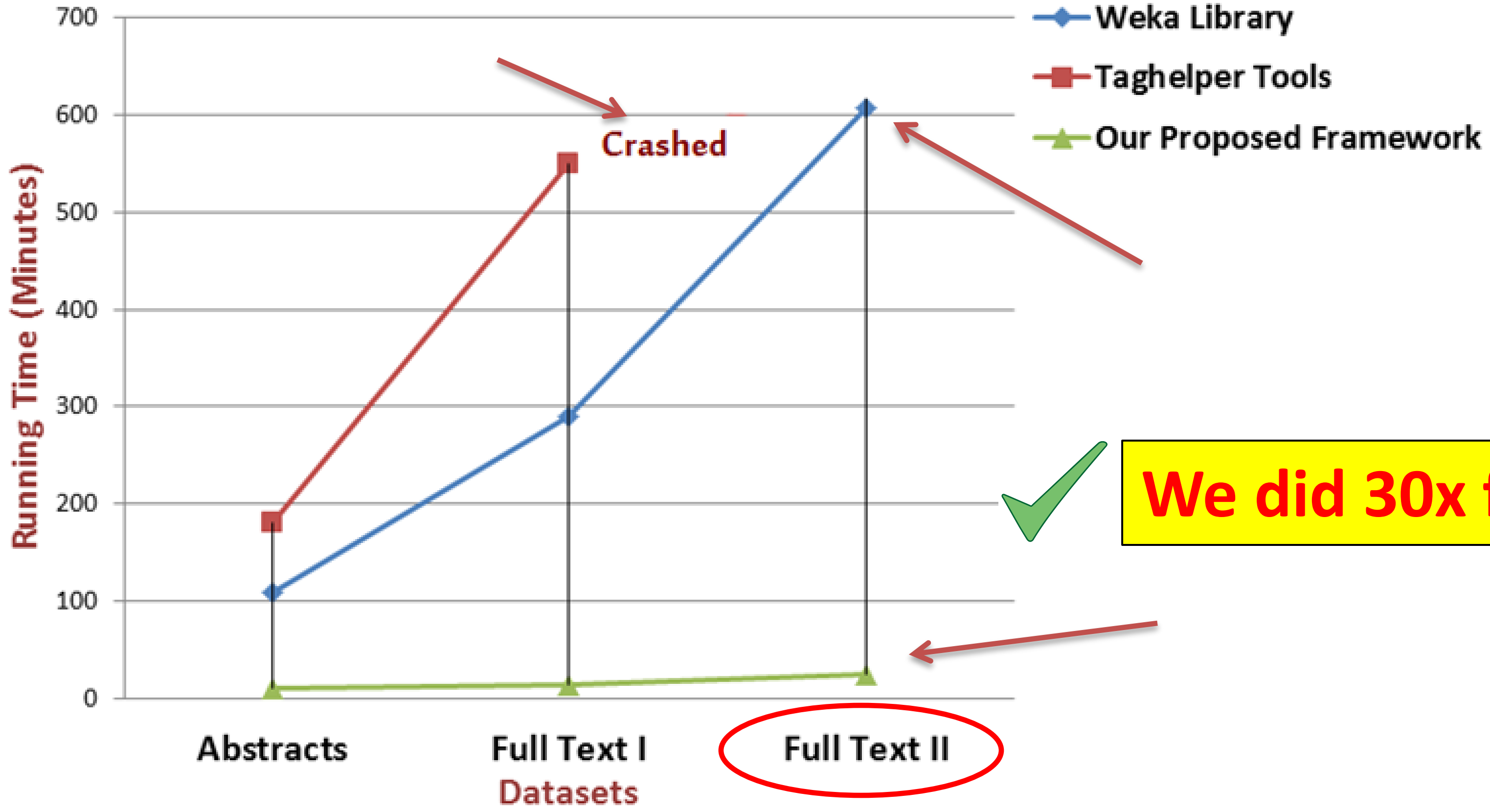


Dataset	Classifier	Accuracy
Abstracts	Naïve Bayes	**96.6%**
Full Text I	Naïve Bayes	**94.8%**
Full Text II	Naïve Bayes	**95.1%**

Results| Summary of the Accuracy

- The accuracy of the prediction model using abstracts is better than full text articles since abstracts have less relevant features.
- Concerning the dataset we have, The accuracy of **Naïve Bayes** classification algorithm is better than SVM, and Logistic Regression.
- Comparing to the well-known **Weka Library** (http://www.cs.waikato.ac.nz/ml/weka/) and **Taghelper Tools** (http://www.cs.cmu.edu/~cprose/TagHelper.html), our proposed framework generates promising results.

Results | The Time Efficiency of the Proposed Framework



We did 30x faster!

* Our proposed framework was developed on a *Big Data* infrastructure.

Conclusions

- Using Big Data infrastructures, we developed a scalable framework for large scale biomedical text mining.
- We worked on almost **16000** abstracts as well as **27000** full text scientific articles published for **cancer research**.
- We worked on ***Big Data*** generated in biomedical research.
- We obtained about **96.6%** accuracy using abstracts, and **94.8%** accuracy using full text articles.
- We did **30x faster** compared with two other tool sets.

When we think together, when we work together, we are bigger, and faster than cancer!

Future Directions

- Employing larger datasets (e.g., 100, 000 or more full text articles).
- Trying dimension reduction on the features extracted from the dataset.
- Developing a comprehensive tool set as a reusable and platform independent software to automatically perform large scale text classification for cancer research.
- We will make it freely available for the research community.

Acknowledgements

- **Center for Human Genetics (CHG):**
  - Max He, Ph.D.
  - Murray Brilliant, Ph.D.
  - Marlene Stueland
  - Cathy Marx

- **Program Directors:**
  - Bobbi Bradley, MPH
  - Huong McLean, Ph.D.

- **Marshfield Clinic Research Foundation, SSRIP, and the donors**.

Thank you very much!

Question?