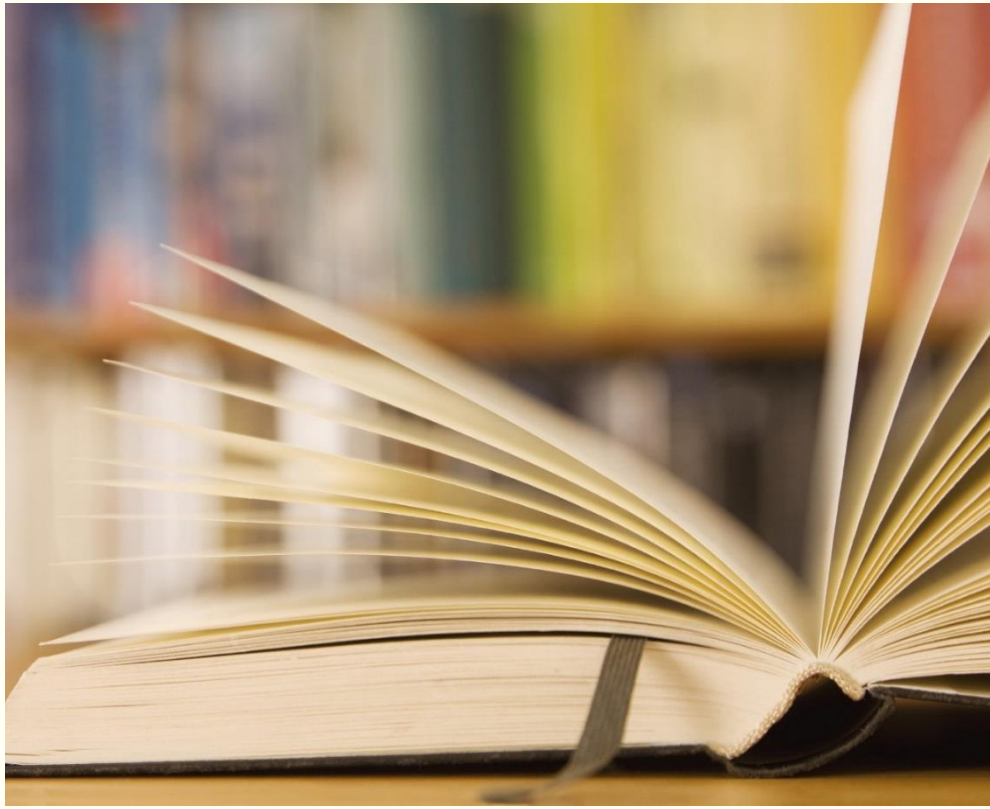
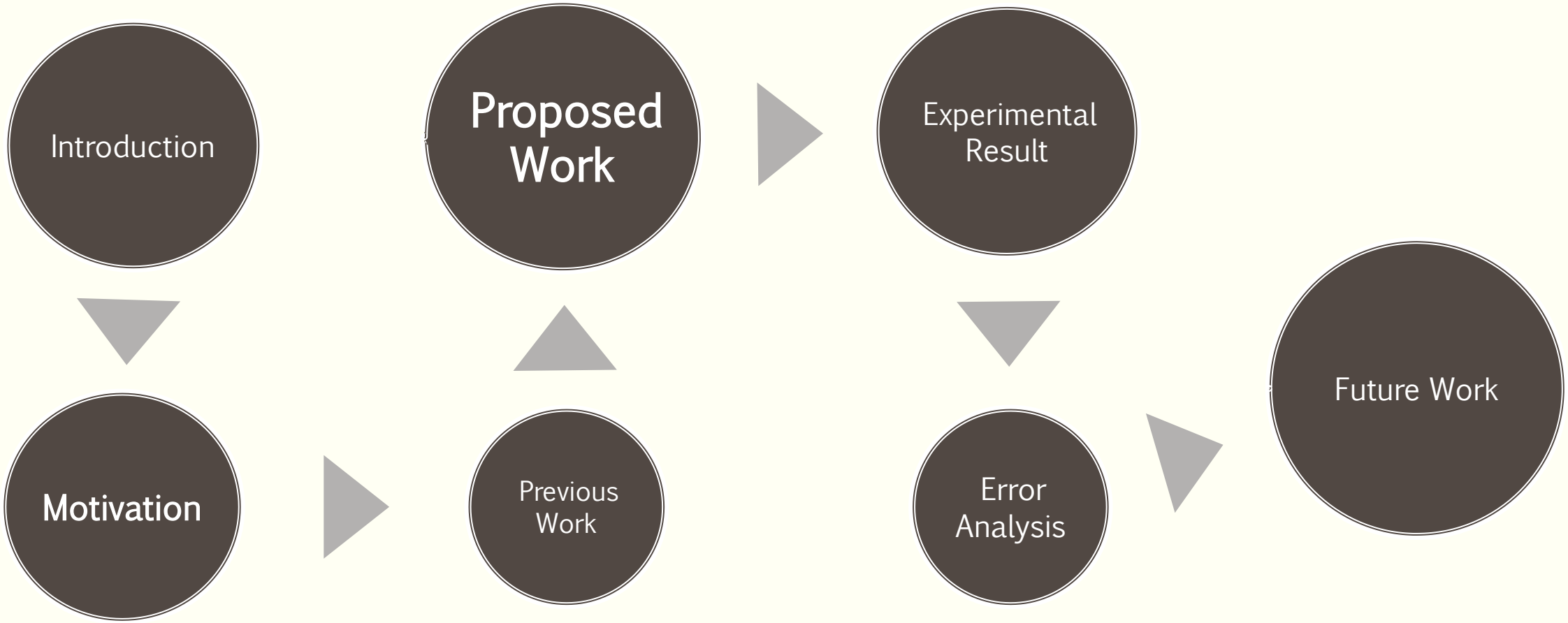
AUTOMATED SEGMENTATION AND CLASSIFICATION OF CHEMICAL AND OTHER EQUATIONS FROM DOCUMENT IMAGES

Prerana Jana, Anubhab Majumdar,
Ashish Kumar Layek, Sekhar Mandal, Amit Kumar Das

Computer Science & Technology Department
IIEST Shibpur, India



Overview


Introduction
Motivation
Previous Work
Proposed Work
Experimental Result
Error Analysis
Future Work

Introduction

Mathematical Equations Segmentation Is Possible!

innovating bank permission to enter new markets into which the technology has disseminated.[11] For simplicity, assume that the technology leader faces only one other bank in each market and that all competitors acquire the screening technology in period 2. At that time, bank 1 obtains permission to enter $n \geq 0$ new markets in addition to its original one so that, with a slight abuse of notation, its *ex ante* expected net profits are now

$$E\left[\Pi_1\left(n\right)\right] = \frac{1}{2}\left(2\phi - 1\right) - \left(\phi - \frac{1}{2}\right)^2 + \left(n+1\right)\phi(1-\phi)(\bar{p}R - 1)$$

Maximization with respect to ϕ yields

$$\phi_n^* = \frac{\left(n+1\right)\left(\bar{p}R - 1\right) + 2}{2 + 2\left(n+1\right)\left(\bar{p}R - 1\right)}$$

Simple differentiation of ϕ_n^* with respect to the number of additional market now yields

$$\frac{\partial \phi_n^*}{\partial n} = -\frac{\bar{p}R - 1}{2\left[1 + \left(n+1\right)\left(\bar{p}R - 1\right)\right]^2} < 0$$

Hence, allowing the technology leader to enter new markets reduces the level of financial innovation if regulators cannot commit to a sufficiently low rate of technological diffusion.

This somewhat surprising result again highlights the trade-off associated with the diffusion of technological progress. When technology developed by one bank becomes available to competitors, two effects operate. First, an increase in ϕ leads to more efficient screening, which increases banks' profits. Second, with greater ϕ banks are subject to lower adverse selection problems and therefore compete more aggressively for borrowers, thus reducing all banks' profits. Which effect dominates then depends on the strategic benefits of screening potential borrowers. If the technology leader

But what about chemical equations ???

- Chemical equations share similar spatial properties as mathematical equations

$$AgCl + 2\ NH_3 \rightarrow Ag(NH_3)_2Cl$$

$$CuSO_4 + 5\ H_2O \rightarrow CuSO_4 \cdot 5H_2O$$

$$\varphi = \frac{b}{a} = \frac{a+b}{b} = (1+\sqrt{5})/2$$

$$H_3O^+ + NH_3 = H_2O + NH_4^+$$

Motivation

- Segment out chemical equations from document images having both chemical and other equations
- Improvement of OCR performance
- Creation of chemical database in searchable format
- Creation of bond electron matrix
- Balancing of chemical equations

Previous Work

- This study is a first of its kind as far as our knowledge goes
- Most equation segmentation work has been targeted towards mathematical equations
- Few techniques applied for the purpose :
    - Symbol Recognition
    - Character size and font information with Bag of word model – math and text bag
    - Check text style (italic, bold or regular) at the character level
- All the above mentioned procedures can’t differentiate mathematical equations from chemical equations

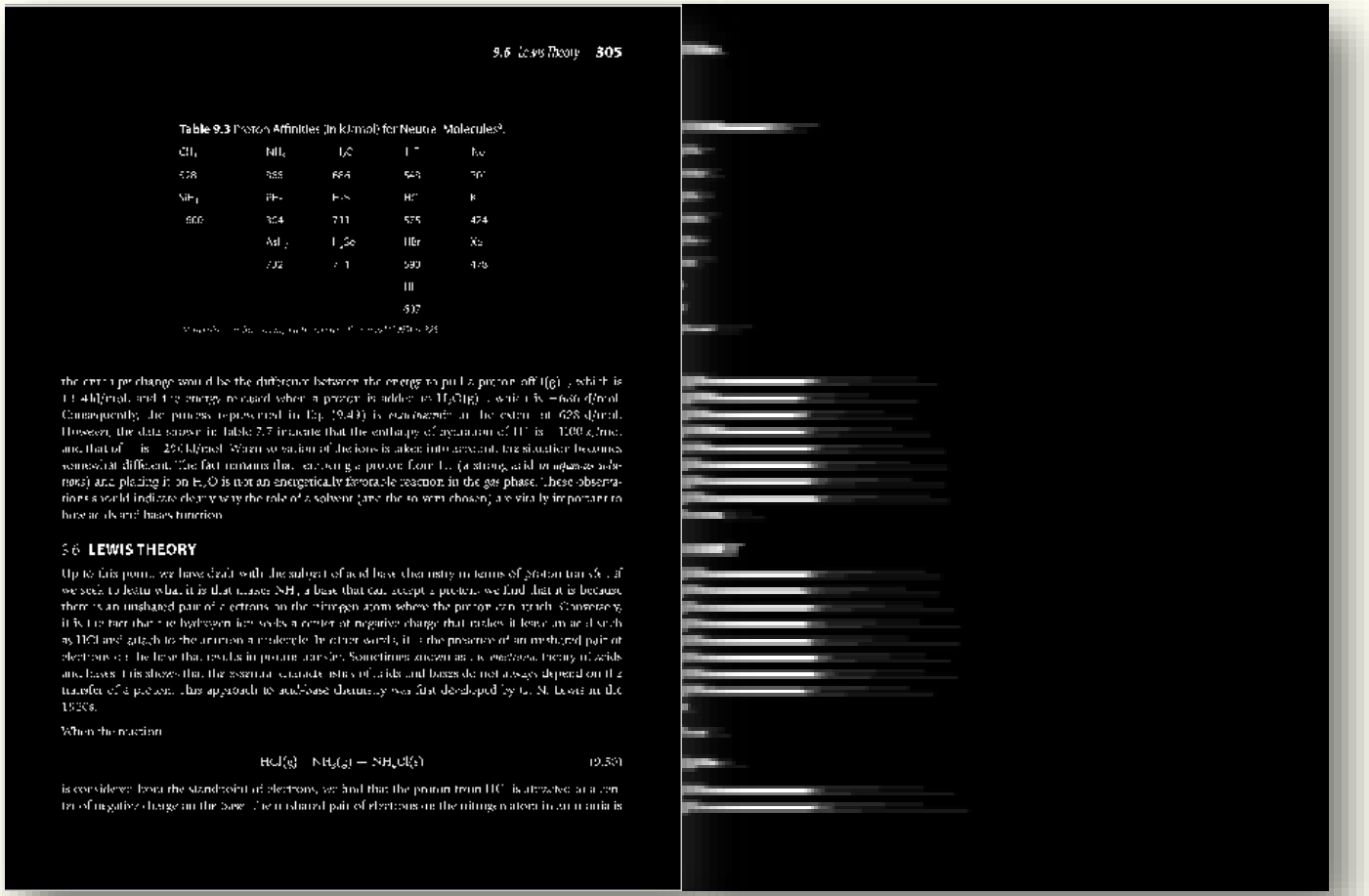
Proposed Work

1. · Text Line Segmentation
2. · Blob formation using morphological tools
3. · Operator Identification
4. · Displayed Equation(DE) zone segmentation
5. · Classification of DE zones into chemical and other zones

Pre-Processing

- Remove small components like
  - Dots of ‘i’, ‘j’
  - Small punctuations like , ; .
- This is done by component labelling and using a suitable threshold

Text Line Segmentation



Text document

Horizontal Projection Profile

Blob Formation using Morphological Operations

$$NaH + H_2O \rightarrow H_2 + NaOH \qquad (10.1)$$

could be presumed to take place in water that exists as H^+OH^-, it is by no means the case. In fact, even though water undergoes a *slight* degree of autoionization, the reaction takes place as shown above

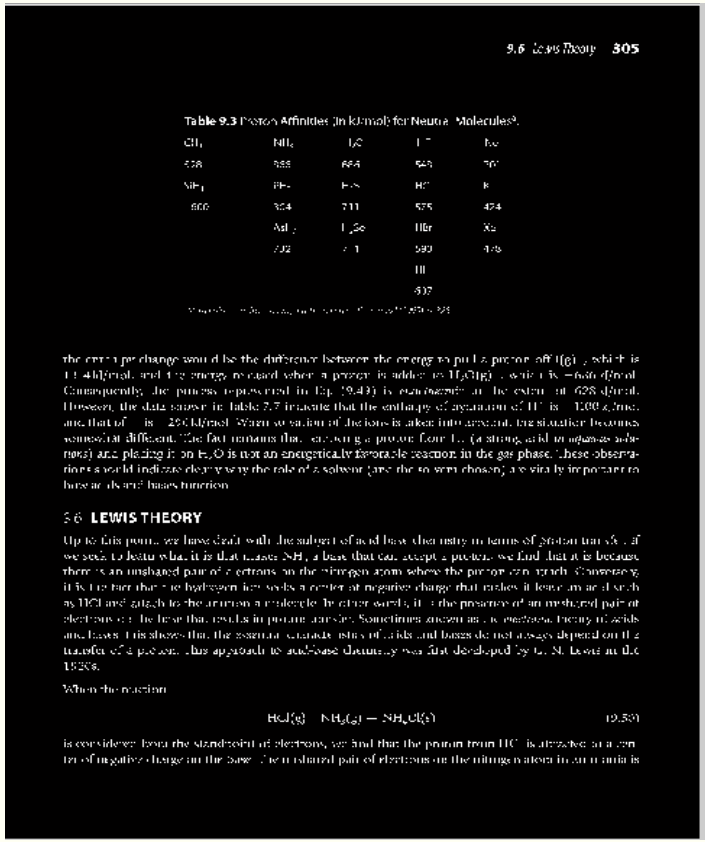
$$NaH + H_2O \rightarrow H_2 + NaOH \qquad (10.1)$$

could be presumed to take place in water that exists as H^+OH^-, it is by no means the case. In fact, even though water undergoes a *slight* degree of autoionization, the reaction takes place as shown above

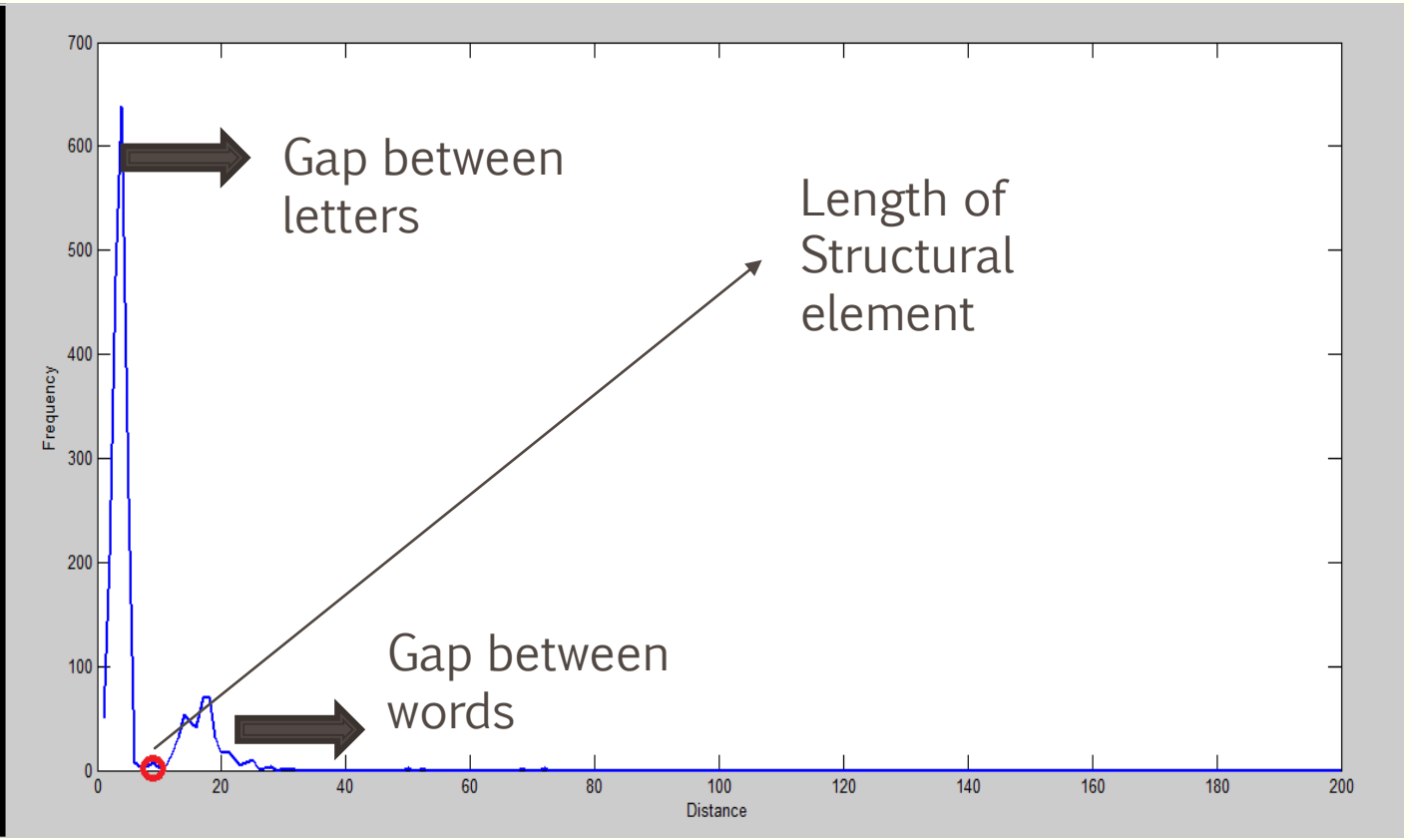
- Structural element used is a **LINE**
- How to determine its length ?

Determining Gap between Characters and Word

- Length of structural element should be greater than the gap between 2 letters in a word but less than the gap between 2 words



Document image



Histogram of character gap

Operator Identification – Single Character Extraction

he solvent is the acidic species and the anion characteristic of the solvent is the basic species. This is known as *the solvent concept*. Neutralization can be considered as the reaction of the cation and anion from the solvent. For example, the cation and anion react to produce unionized solvent.

$$HCl + NaOH \rightleftharpoons NaCl + H_2O \quad (10.4)$$

$$NH_4Cl + NaNH_2 \rightarrow NaCl + 2\,NH_3 \quad (10.5)$$

Note that there is no requirement that the solvent actually undergo autoionization.

$$+ \rightleftharpoons + \quad (4)$$

$$+ \rightarrow + 2 \quad (5)$$

Operator Identification – Euler Number based removal

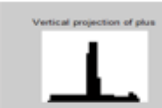
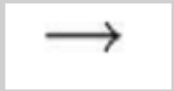
$$+ \quad \rightleftharpoons \quad + \qquad (4)$$

$$+ \quad \rightarrow \quad +2 \qquad (5)$$

$$+ \quad \rightleftharpoons \quad + \qquad (\ )$$

$$+ \quad \rightarrow \quad +2 \qquad (5)$$

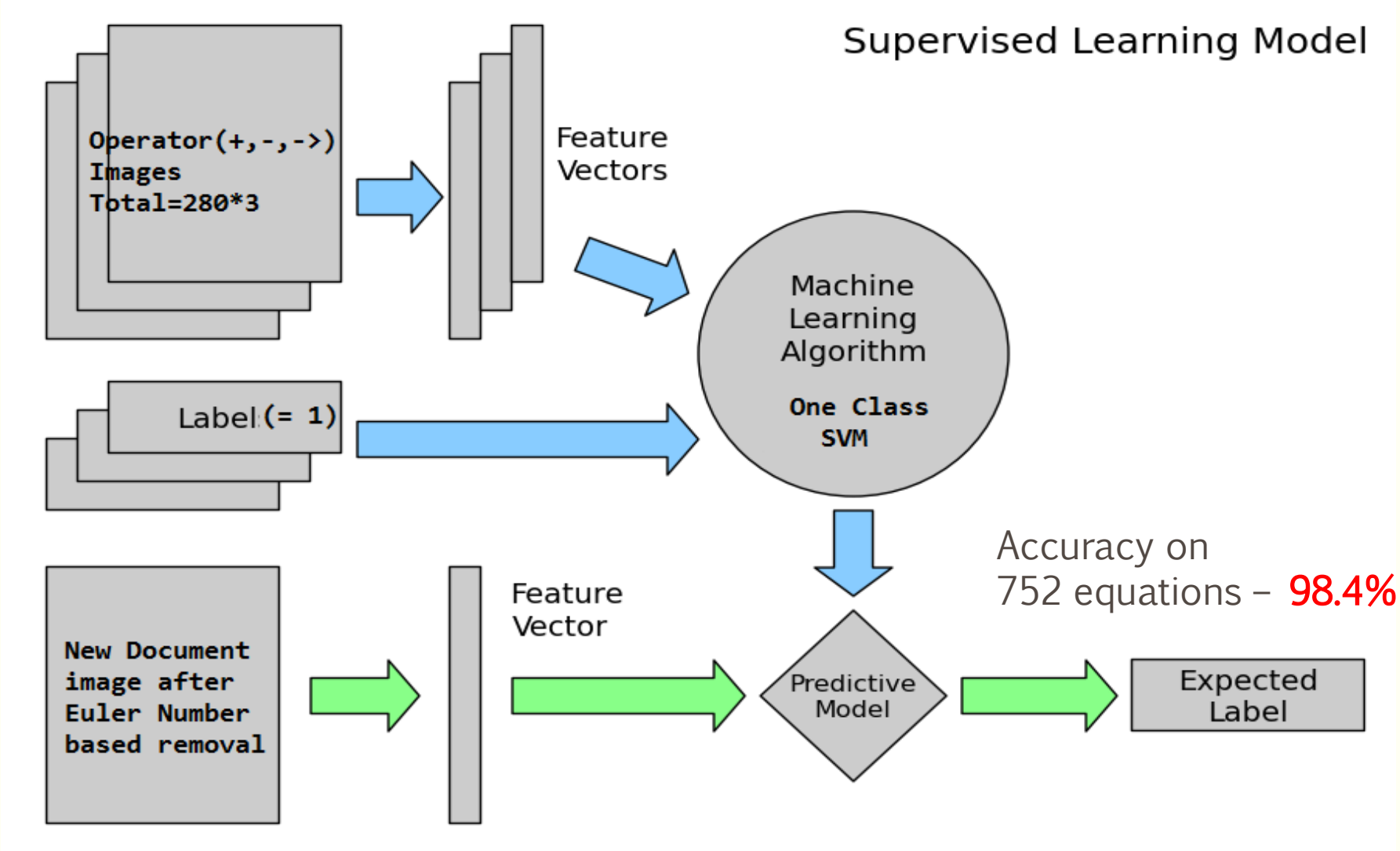
Operator Identification – Feature Extraction

Operators	Horizontal Projection	Vertical Projection
		
		

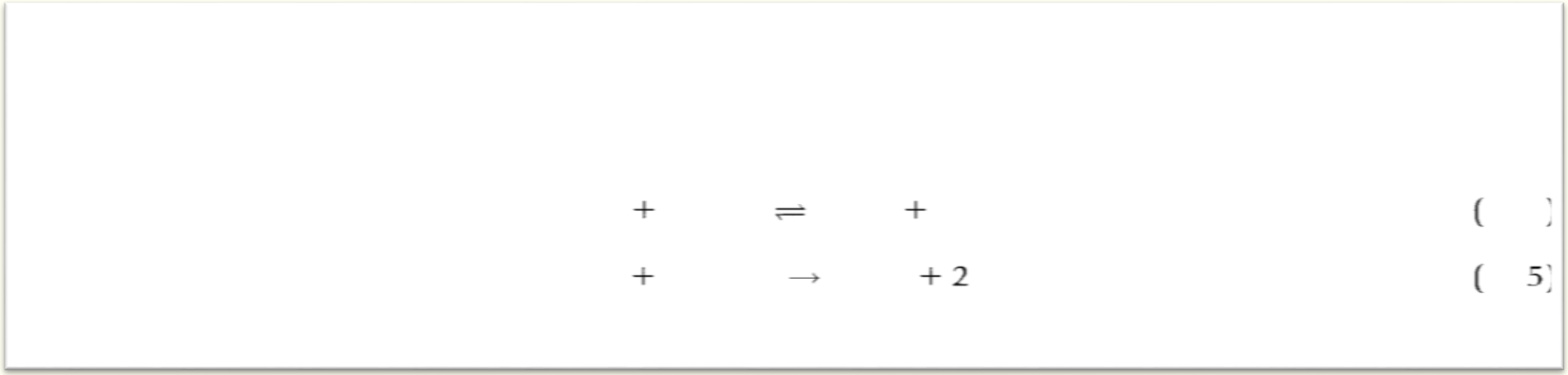
The following features are considered for remaining single characters :

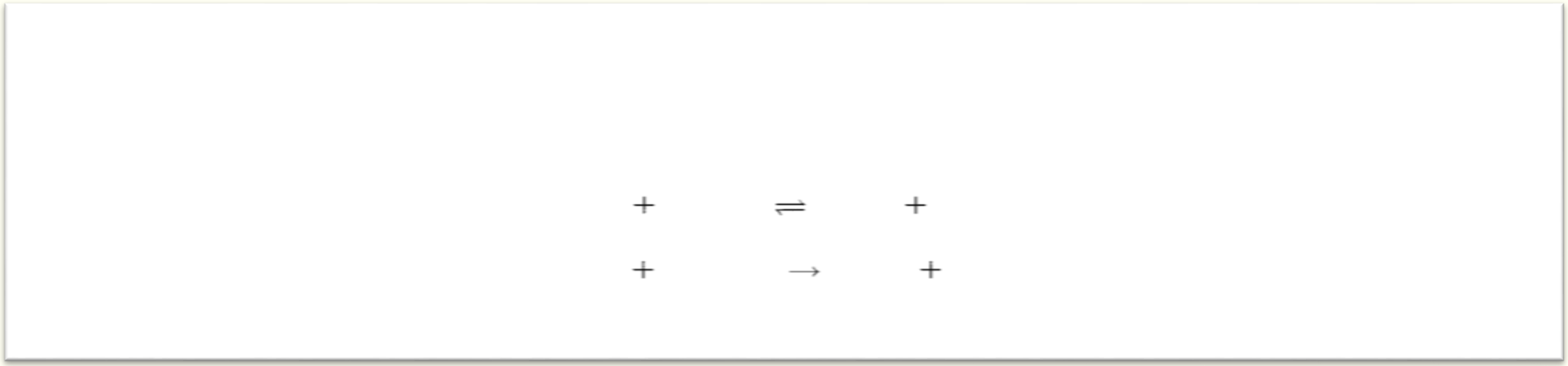
- Aspect ratio of each component
- Density (#object pixel/#total pixel)
- Second and third order moments of horizontal and vertical projection profile
- The location and magnitude of global maxima in horizontal and vertical projection profile
- Perimeter of the component

Operator Identification – One Class SVM



After Operator Identification - sample output





Identifying '=' or '⇌'

- '=' or '⇌' doesn't form a word blob
- The upper and lower part is identified as thin character : '-' or '→'
- For every thin character detection we place a mask of size (l*l/2) on top and bottom to check presence of another such symbol where l=length of the thin operator.
- If present, we join them together and it forms one single operator.

Adding numerator and denominator

Original Equation	$\gamma = \frac{}{R_g T^0}$
Lower Scan	$R_g T^0$
Upper Scan	E
Final Equation	$\gamma = \frac{E}{R_g T^0}$

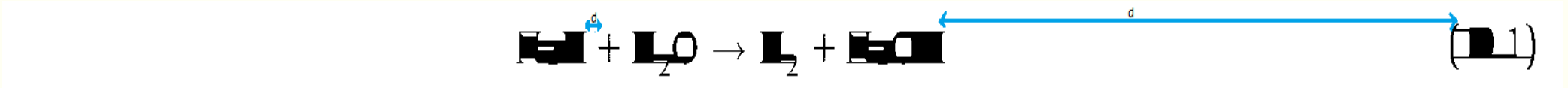
Segmentation of DE zones - Run Length Smoothing on Operands

$$NaH + H_2O \rightarrow H_2 + NaOH \qquad (10.1)$$

$$NaH \quad H_2O \quad H_2 \quad NaOH \qquad (10.1)$$

$$[redacted] \qquad (10.1)$$

Segmentation of DE zones - Equation Number Removal



$$NaH + H_2O \rightarrow H_2 + NaOH$$

Segmentation of DE zones – Displayed or Embedded?

- Count operands in CDE

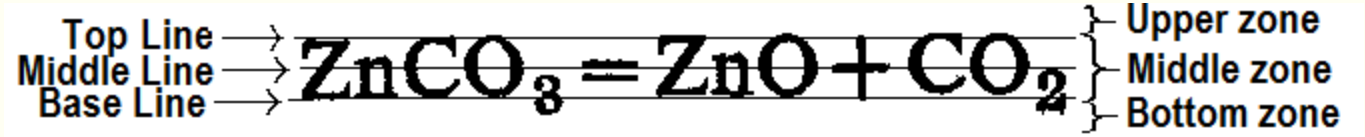
- Count operators on CDE

- A CDE is considered as a DE if

$$\#\text{Operands} \leq 2*(\#\text{Operators})$$

This is because any operator will have at most 2 operators on its either side

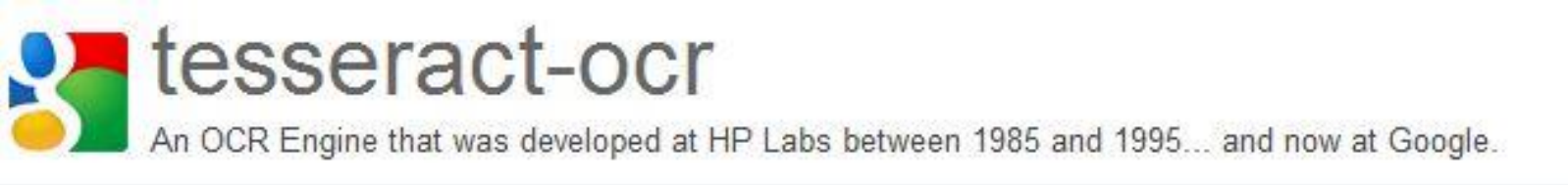
Classification of Segmented DE zones – Removal of Subscripts and Superscripts



- Subscripts Ɛ (Lower middle zone, Bottom Zone)
- Superscripts Ɛ (Upper middle zone, Upper zone)
- Chemical elements won't be subscript or superscript; so they are ignored

Classification of Segmented DE zones – OCR

- Rest are input into Google Tesseract 3.02, an open-source OCR engine, which returns the text form.



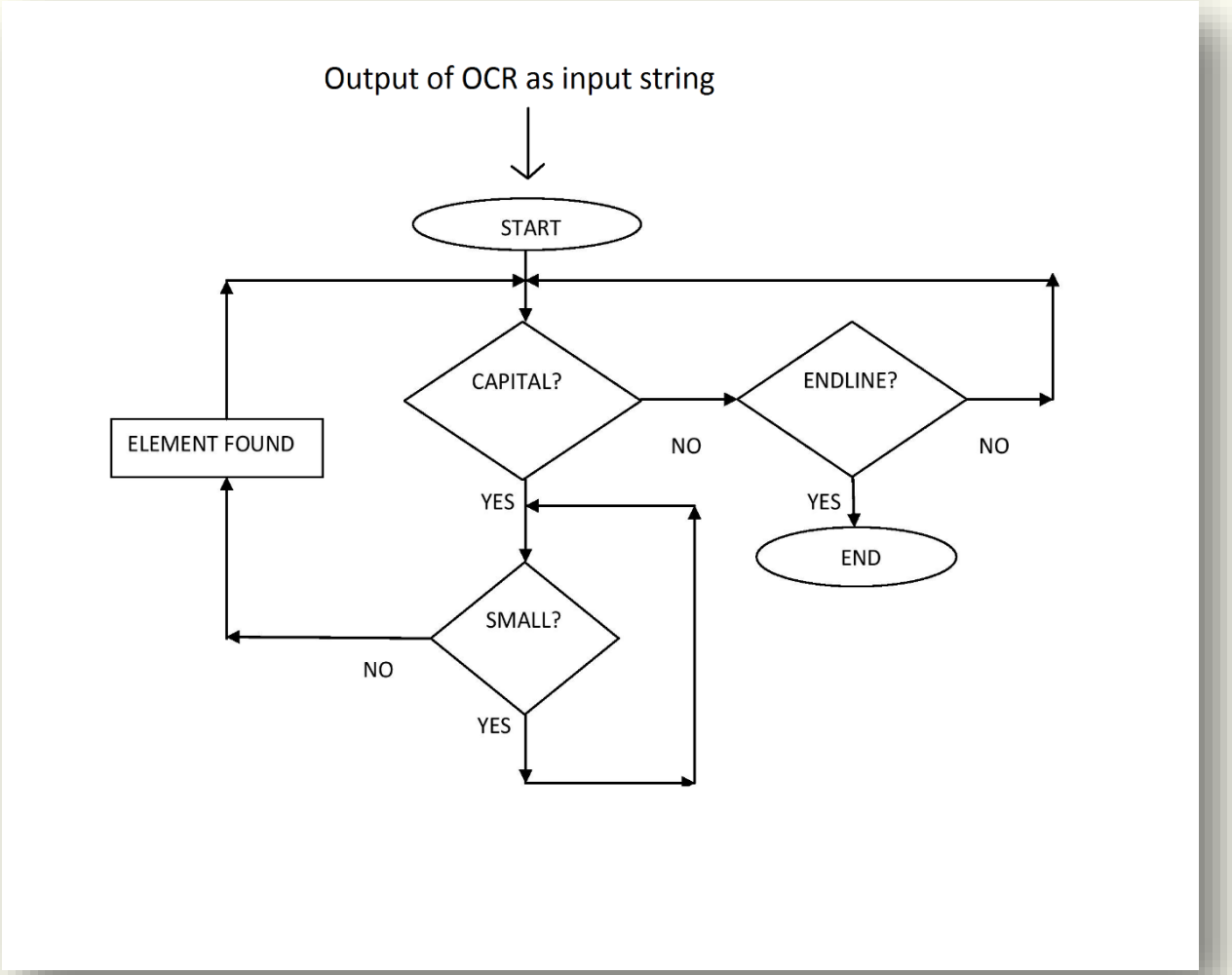
Classification of Segmented DE zones – Parser

- Regular expression for a periodic element

→ [A-Z][a-z]*

- Grammar for substrings that can be an element:
  - start → capital . follow
  - follow → small . follow | Ɛ
  - capital → A|B|C|...|X|Y|Z
  - small → a|b|c|...|x|y|z

Flowchart of the working mechanism of the parser



Classification of Segmented DE zones – Chemical or Other?

- The substrings are matched against a hash table consisting of all the elements in the Periodic Table
- If #(elements matched) : #(total substrings) ≥ 0.7 then the DE is considered as chemical equation (according to our experiment with 733 equations)

- Why 0.7, not 1?

→Limitation of OCR

→Presence of broken and touching character in DE zone

Experimental Result

- Algorithm implemented on MATLAB R2014a
- Total document images → 152
- Total displayed equations → 752

Actual \ Classified As	Chemical	Other
Chemical	97.8%	2.2%
Other	2.95%	97.05%

Error Analysis – I. Segmentation Error

Case 1:

$$Na_2SO_4 \xrightarrow{H_2O} Na^+ + SO_4^{2-}$$

Case 2:

Proof of Proposition 3. Maximization of bank 1's *ex ante* expected net profits in Equation (4), $E[\Pi_1] = (2-\lambda)\frac{1}{2}(2\phi-1) - \left(\phi-\frac{1}{2}\right)^2 + \lambda\phi(1-\phi)(\bar{p}R-1)$, with respect to ϕ yields $\phi^* = \frac{3+\lambda(\bar{p}R-2)}{2+2\lambda(\bar{p}R-1)}$ from the FOC

$$\frac{\partial E[\Pi_1]}{\partial \phi} = 3 - \lambda + \lambda(1-2\phi)(\bar{p}R-1) - 2\phi = 0$$

Since $\lambda \in [0,1]$, we can always choose $\bar{p}R < M$, for sufficiently small M, such that the optimal

Error Analysis – II. Classification Error

12.1.1 Reaction of Metals and Alkyl Halides

This technique is most appropriate when the metal is highly reactive. It should be kept in mind that even though a formula may be written as if the species is a monomer, several types of organometallic compounds are associated. Examples of this type of reaction are

$$2\ Li + C_4H_9Cl \rightarrow LiC_4H_9 + LiCl \tag{12.1}$$

$$4\ Al + 6\ C_2H_5Cl \rightarrow [Al(C_2H_5)_3]_2 + 2\ AlCl_3 \tag{12.2}$$

$$2\ Na + C_6H_5Cl \rightarrow NaC_6H_5 + NaCl \tag{12.3}$$

The most important reaction of this type is that in which Grignard reagents are produced:

$$RX + Mg \xrightarrow{\text{dry ether}} RMgX \tag{12.4}$$

Future Work -

- Correction of the segmentation error for reactants over the arrow

- Bond Electron Matrix Formation

  → Recognizing Subscripts and Superscripts

  → Improvement of OCR output

THANK YOU!