

UNIVERSITY OF CALIFORNIA DAVIS  
DEPARTMENT OF STATISTICS

---

# Machine Learning (STA-208)

---

Ying-Chen Chou  
Chia-Hui Shen  
Jiahui Tan  
Pei-Ying Ling



# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Description of Data</b>	<b>2</b>
2.1	Raw Data . . . . .	2
2.2	Top 50 Words Per Year . . . . .	2
2.3	Word Cloud . . . . .	2
2.4	Mathematics . . . . .	2
<b>3</b>	<b>Previous Studies</b>	<b>2</b>
<b>4</b>	<b>Method</b>	<b>2</b>
4.1	Feature Engineering . . . . .	2
4.2	Fit Models . . . . .	2
4.3	Conclusion and exploration . . . . .	2

# **1 Introduction**

## **2 Description of Data**

### **2.1 Raw Data**

Year ( 1999 - 2007 )

Sample ( Total )

Source (Difficulty of cleaning, how to deal with it( combination of different source, HTML...))

### **2.2 Top 50 Words Per Year**

### **2.3 Word Cloud**

### **2.4 Mathematics**

## **3 Previous Studies**

## **4 Method**

### **4.1 Feature Engineering**

### **4.2 Fit Models**

### **4.3 Conclusion and exploration**