Mining-based Software Engineering

CS585 Software Verification and Validation
February 25th, 2015
http://cs585.yusun.io

Dr. Yu Sun
http://yusun.io
yusun@cpp.edu

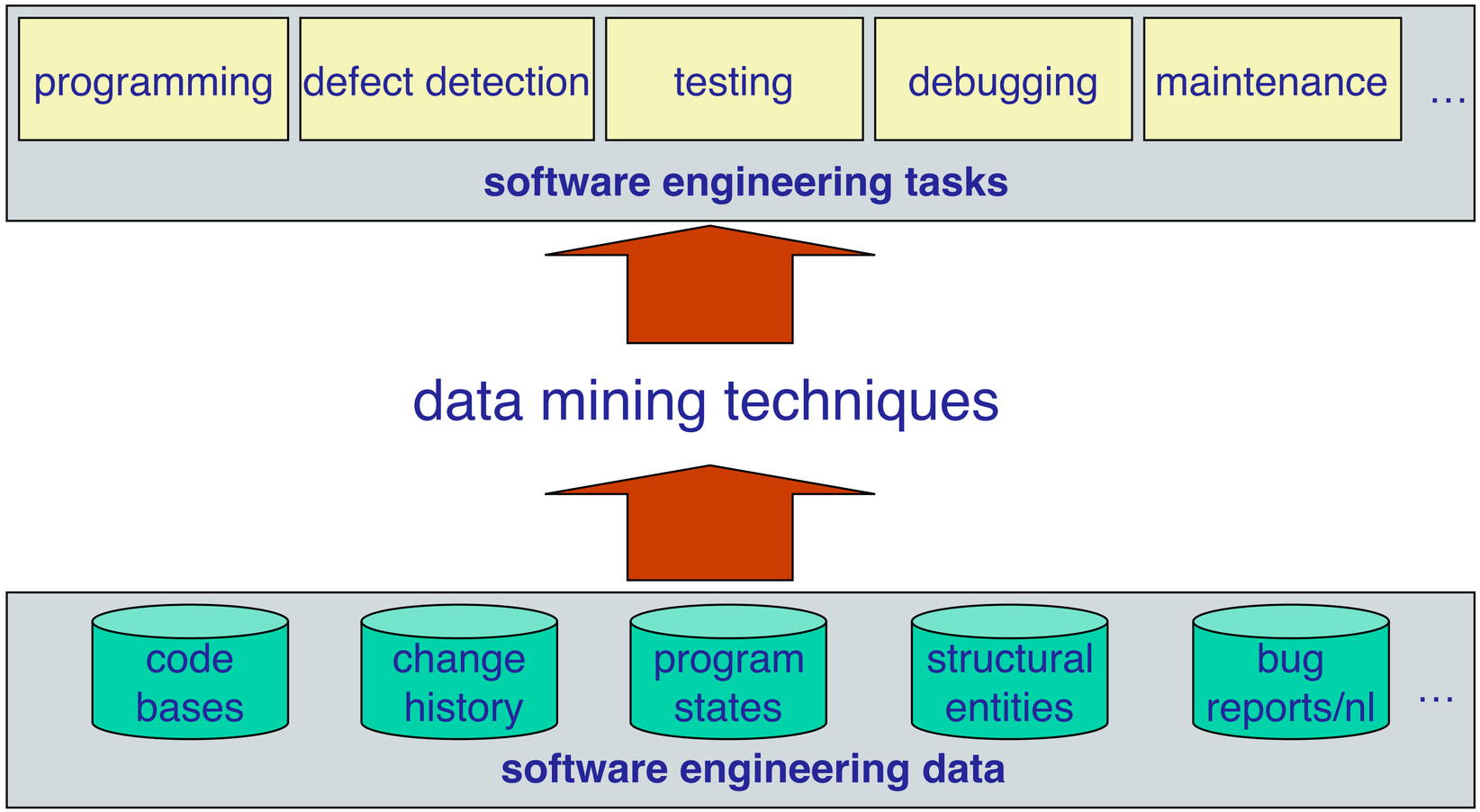
CAL POLY POMONA

Data mining is the process to extract information from a data set and transform it into an understandable structure for further use.

+

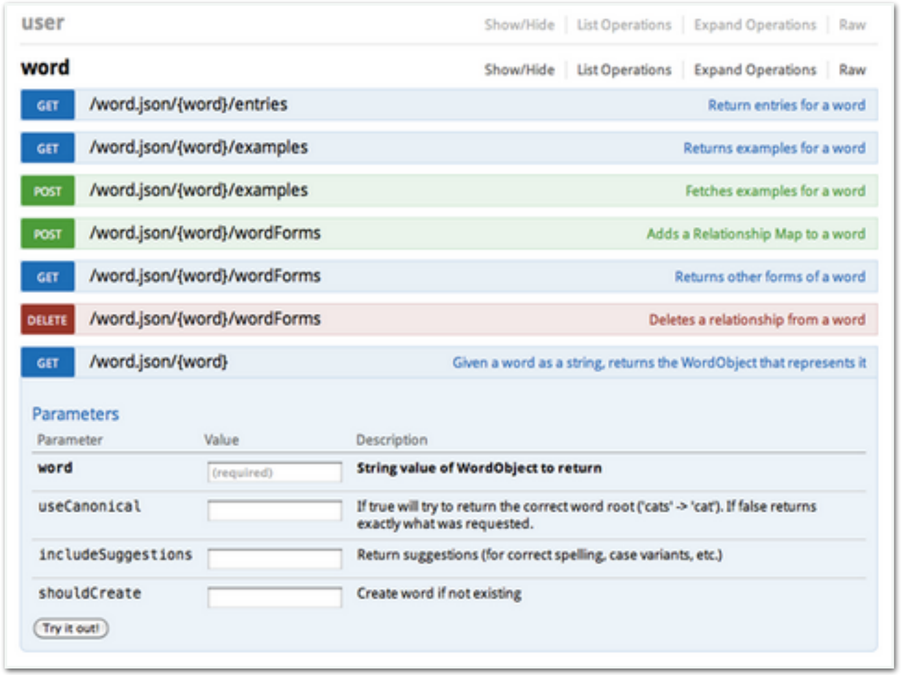
Big data is used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

What software engineering tasks can be helped by data mining?



Mining API Patterns

- Developers reuse existing APIs
- Complexity and lack of documentation
- Spend more efforts in understanding APIs
- Introduce defects in API client code
- Mining API properties as common patterns across API client code



Acharya, Mithun, Tao Xie, Jian Pei, and Jun Xu. "Mining API patterns as partial orders from source code: from usage scenarios to specifications." ECOOP 2007

Finding Copy-Paste Bugs

- Copy-paste is bad
- Finding copy-paste is easy, but detecting is difficult
- Copy-Paste bug detection using static analysis

```
( linux-2.6.6/arch/sparc64/prom/memory.c )
68 void __init prom_meminit(void)
69 {
        ......
92     for(iter=0; iter<num_regs; iter++) {
93       prom_phys_total[iter].start_adr =
94          prom_reg_memlist[iter].phys_addr;
95       prom_phys_total[iter].num_bytes =
96          prom_reg_memlist[iter].reg_size;
97       prom_phys_total[iter].theres_more =
98          &prom_phys_total[iter+1];
99     }
        ......
111    for(iter=0; iter<num_regs; iter++) {
112      prom_prom_taken[iter].start_adr =
113         prom_reg_memlist[iter].phys_addr;
114      prom_prom_taken[iter].num_bytes =
115         prom_reg_memlist[iter].reg_size;
116      prom_prom_taken[iter].theres_more =
117         &prom_phys_total[iter+1];      // bug
118    }
        ......
143 }
```

CP-Miner: A Tool for Finding Copy-paste and Related Bugs in Operating System Code. Zhenmin Li, Shan Lu, Suvda Myagmar, and Yuanyuan Zhou. OSDI 2004.

Discovering Algebraic Specifications from Java Classes

- Formal program specification is useful
- Developing formal specification is hard
- Automated algebraic specification generation

```
public class IntStack {
   private int [] store;    private int size;
   private static final int INITIAL_CAPACITY=10;
   public IntStack(){
      this.store = new int[INITIAL_CAPACITY];
      this.size=0;
   }
   public void push(int value){
      if(this.size ==this.store.length){
         int [] store = new int[this.store.length*2];
         System.arraycopy(this.store,0,store,0,this.size);
         this.store = store;
      }
      this.store[this.size]=value;
      this.size++;
   }
   public int pop(){
      int result = this.store[this.size-1];
      this.size--;
      if(this.store.length > INITIAL_CAPACITY && this.size*2
                < this.store.length){
         int [] store = new int[this.store.length/2];
         System.arraycopy(this.store,0,store,0,this.size);
         this.store = store;
      }
      return result;
   }
}
```

Discovering Algebraic Specifications from Java Classes. Johannes Henkel and Amer Diwan. ECOOP 2003.

Discovering Algebraic Specifications from Java Classes

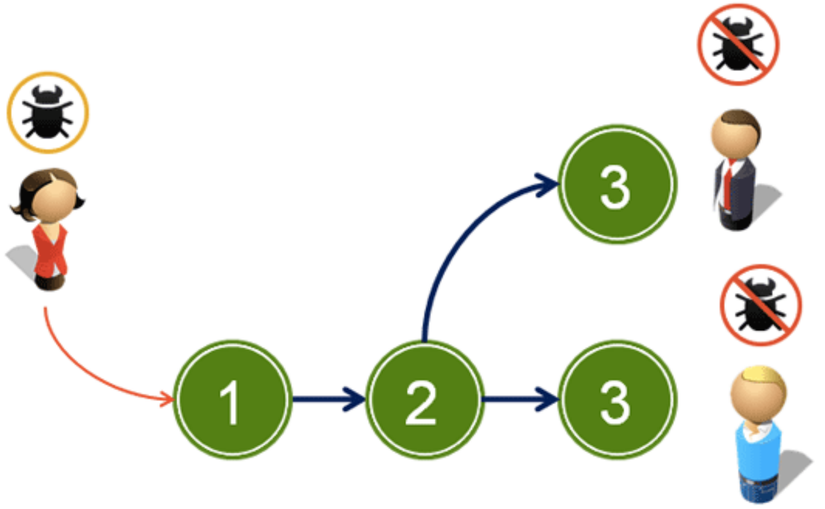
- Formal program specification is useful
- Developing formal specification is hard
- Automated algebraic specification generation

TYPE $IntStack$
FUNCTIONS
$IntStack : \rightarrow IntStack \times void$
$push : IntStack \times int \rightarrow IntStack \times void$
$pop : IntStack \rightarrow IntStack \times int$
AXIOMS
$\forall s : IntStack, i : int$
$pop(push(s, i).state).retval = i$
$pop(push(s, i).state).state = s$
$pop(IntStack().state).retval \rightsquigarrow \text{ArrayIndexOutOfBoundsException}$

Discovering Algebraic Specifications from Java Classes. Johannes Henkel and Amer Diwan. ECOOP 2003.

Who Should Fix This Bug?

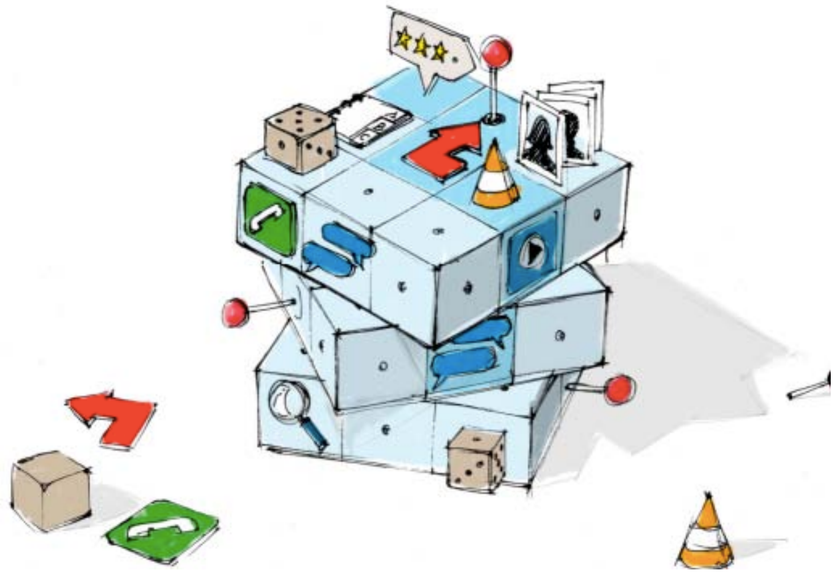
- Assigning bug fix task for large-scale project is challenging
- Learn developers' bug fix history
- Automatically assign tasks



Who Should Fix This Bug? John Anvik, Lyndon Hiew, and Gail C. Murphy. ICSE 2006.

Mining System-User Interaction Logs for Interaction Patterns

- Mine system-level interactions
- Mine user-level interactions
- Help system maintenance
- Help build recommendation system



Mining System-User Interaction Logs for Interaction Patterns. Mohammad El-Ramly and Eleni Stroulia. MSR 2004.

What kinds of software engineering data can be mined?

Static Code Base

Software Change History

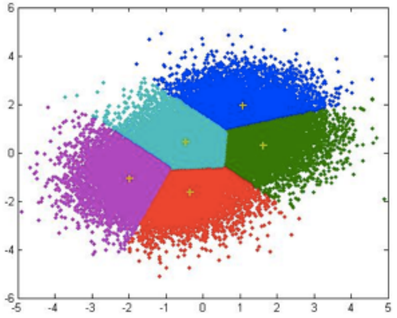
Profiled Structural Entities

Profiled Program States

Bug Reports/Natural Languages

How are data mining techniques used in software engineering?

Classification



Clustering

Automation/Grammar Learning

Search/Matching

Text Mining

Data Mining Made Easy