> ProCode( );

rating source code quality

student professional

Richard Beal

INSIGHT

“We ***all*** write source code.”

“Let’s write ***professional***-quality source code!”

Framework

Language: C

Code:

```
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
```

Framework


```
> ProCode( );
```

Language: C

Code:

```
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
```

Syntax Tokenization: *S*

FLOAT ID3 SEMI
INT ID4 EQ D4 SEMI
ID3 EQ ID4 MULT F7 SEMI

Framework

Language: C

Code:

```
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
```

```
> ProCode( );
```

Syntax Tokenization: *S*

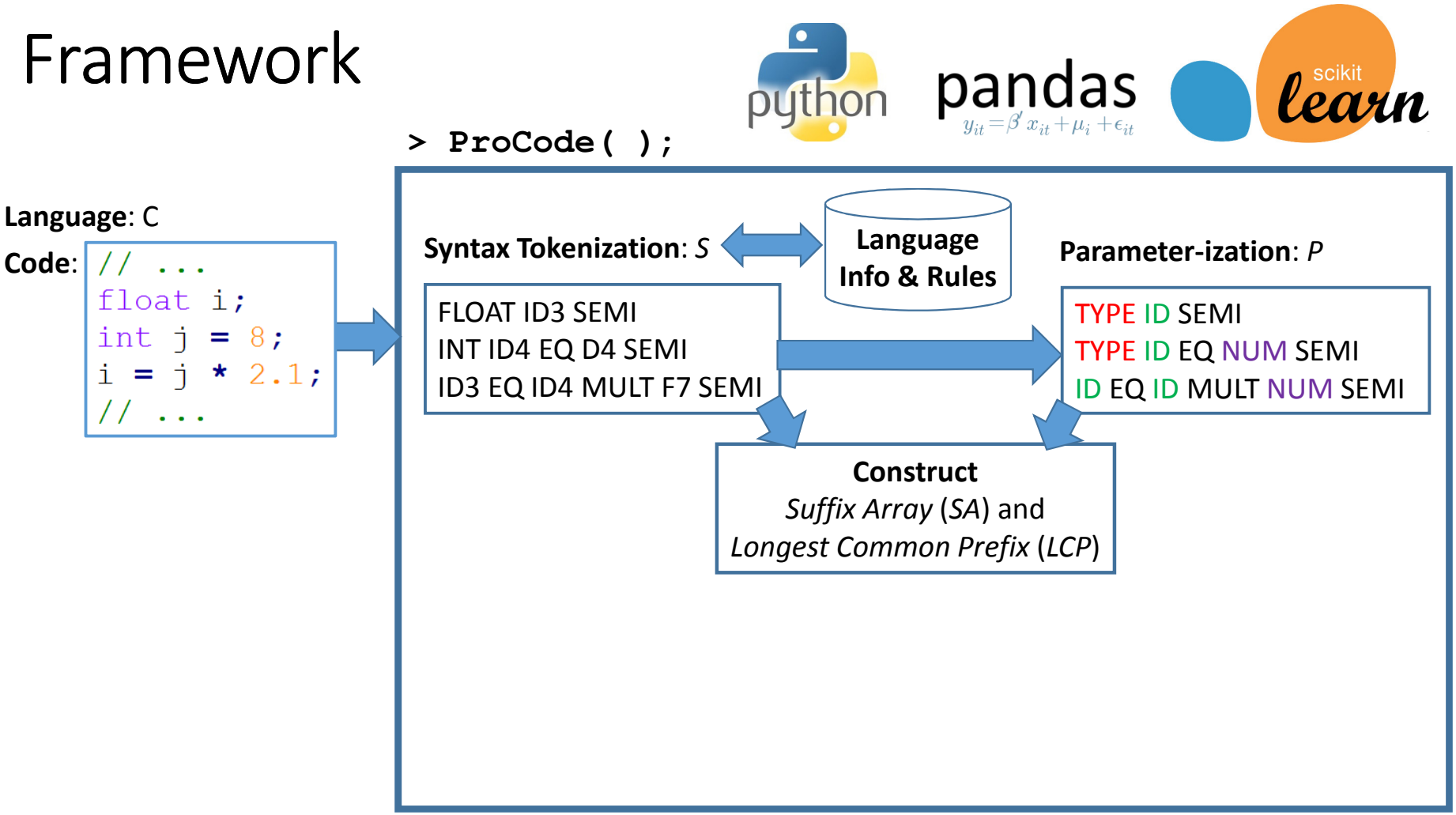
Language Info & Rules

```
FLOAT ID3 SEMI
INT ID4 EQ D4 SEMI
ID3 EQ ID4 MULT F7 SEMI
```

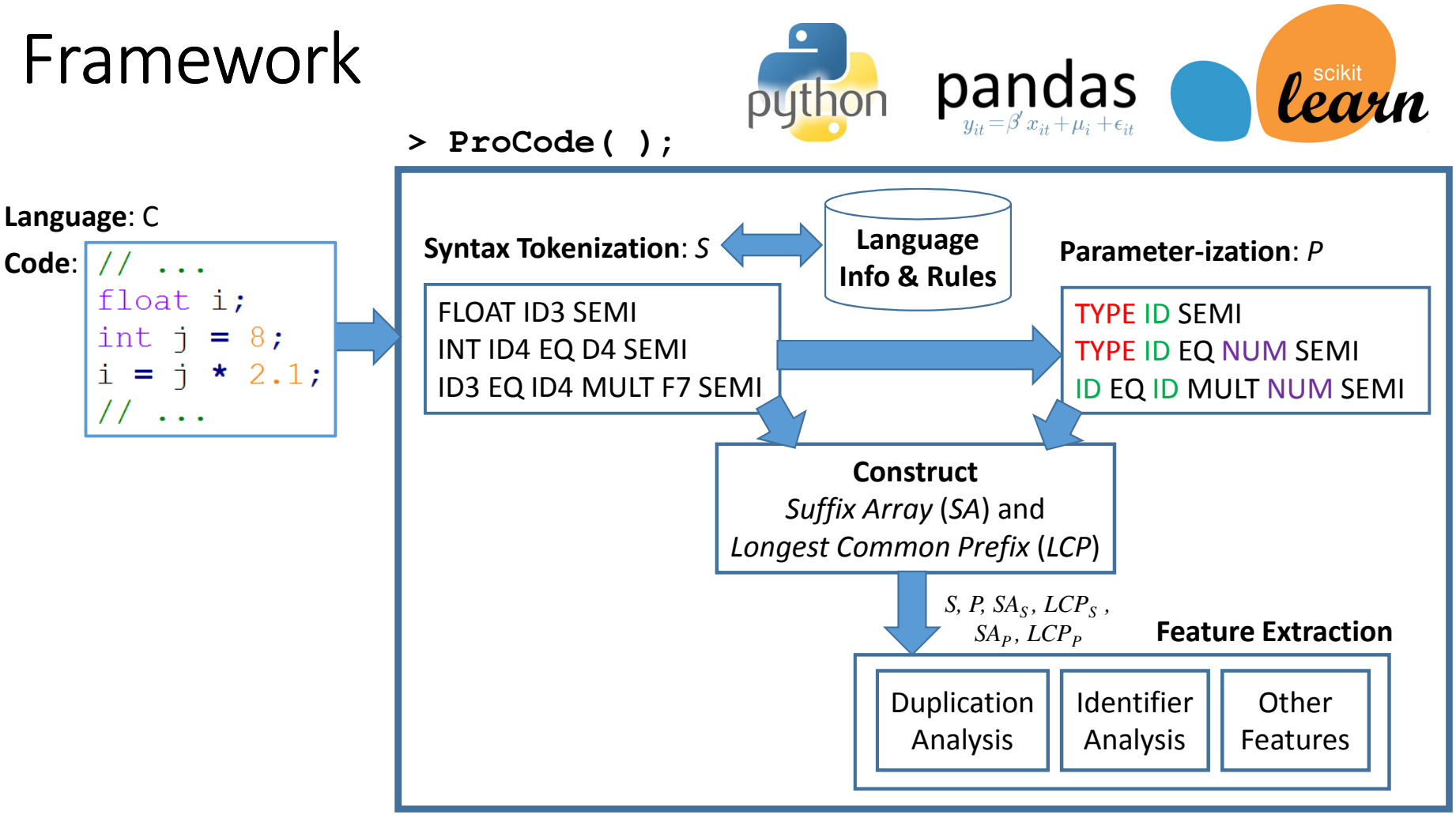
Parameter-ization: *P*

```
TYPE ID SEMI
TYPE ID EQ NUM SEMI
ID EQ ID MULT NUM SEMI
```

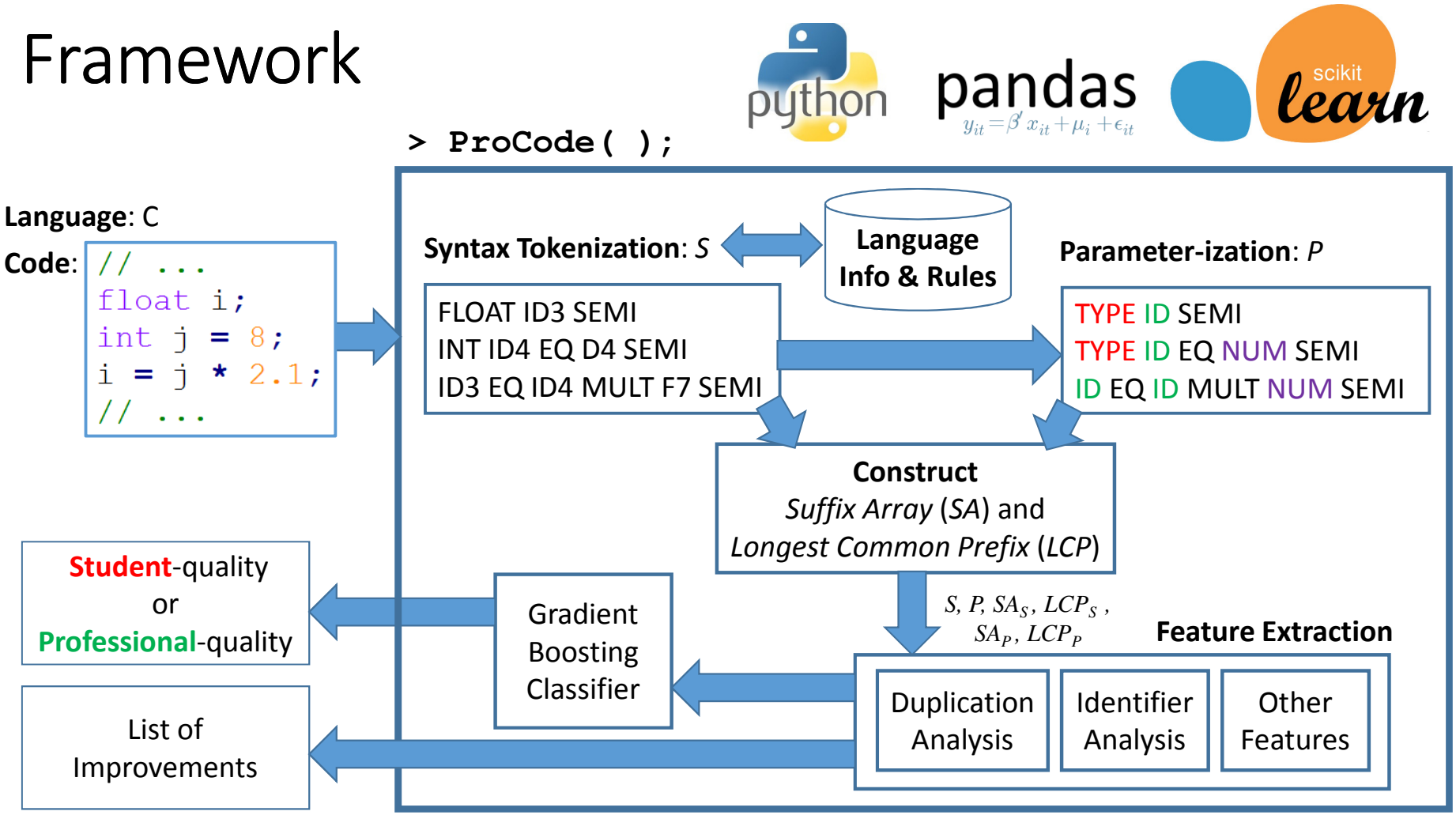
Framework


> ProCode( );
Language: C
Code:
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
Syntax Tokenization: S
Language Info & Rules
Parameter-ization: P
FLOAT ID3 SEMI
INT ID4 EQ D4 SEMI
ID3 EQ ID4 MULT F7 SEMI
TYPE ID SEMI
TYPE ID EQ NUM SEMI
ID EQ ID MULT NUM SEMI
Construct
Suffix Array (SA) and
Longest Common Prefix (LCP)

Framework


> ProCode( );
Language: C
Code:
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
Syntax Tokenization: S
Language Info & Rules
Parameter-ization: P
FLOAT ID3 SEMI
INT ID4 EQ D4 SEMI
ID3 EQ ID4 MULT F7 SEMI
TYPE ID SEMI
TYPE ID EQ NUM SEMI
ID EQ ID MULT NUM SEMI
Construct
Suffix Array (SA) and
Longest Common Prefix (LCP)
S, P, SA_S, LCP_S , SA_P, LCP_P
Feature Extraction
Duplication Analysis
Identifier Analysis
Other Features

Framework

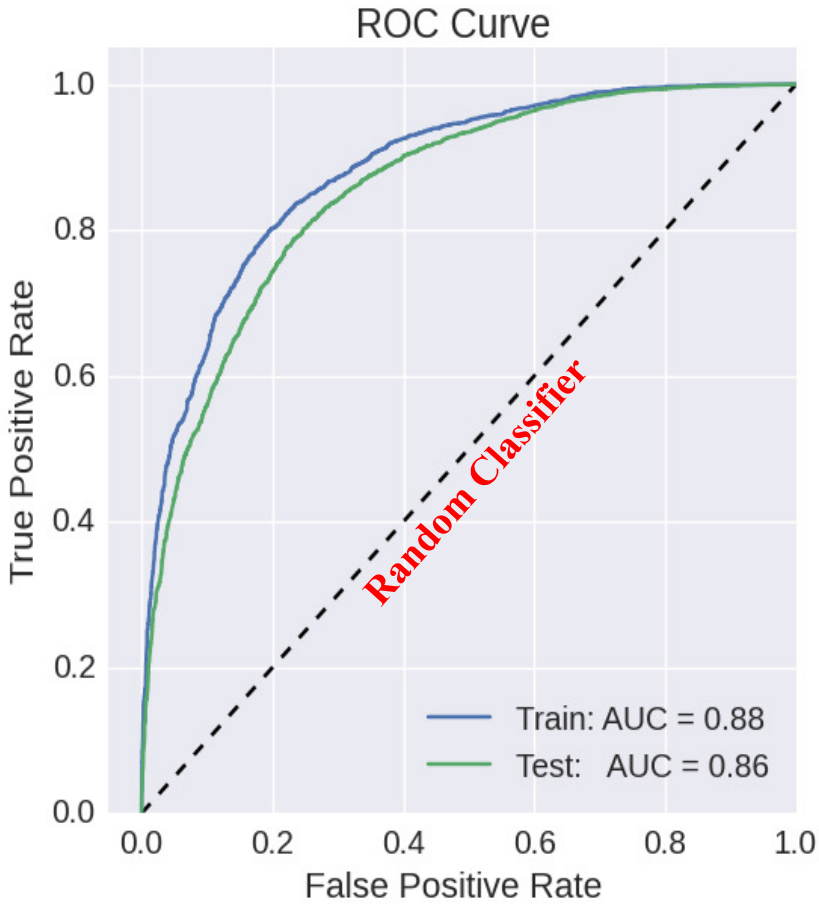

> ProCode( );
Language: C
Code:
// ...
float i;
int j = 8;
i = j * 2.1;
// ...
Syntax Tokenization: S
Language Info & Rules
Parameter-ization: P
FLOAT ID3 SEMI
INT ID4 EQ D4 SEMI
ID3 EQ ID4 MULT F7 SEMI
TYPE ID SEMI
TYPE ID EQ NUM SEMI
ID EQ ID MULT NUM SEMI
Construct
Suffix Array (SA) and
Longest Common Prefix (LCP)
S, P, SA_S, LCP_S , SA_P, LCP_P
Feature Extraction
Duplication Analysis
Identifier Analysis
Other Features
Gradient Boosting Classifier
Student-quality or Professional-quality
List of Improvements

Performance

- Achieved **86%** AUC on labeled C-code from **GitHub**
  - "Professional" = Top-rated repository
  - "Student" = Academic repository
- Classifier uses "good habit" features
  - **Creative** variable names

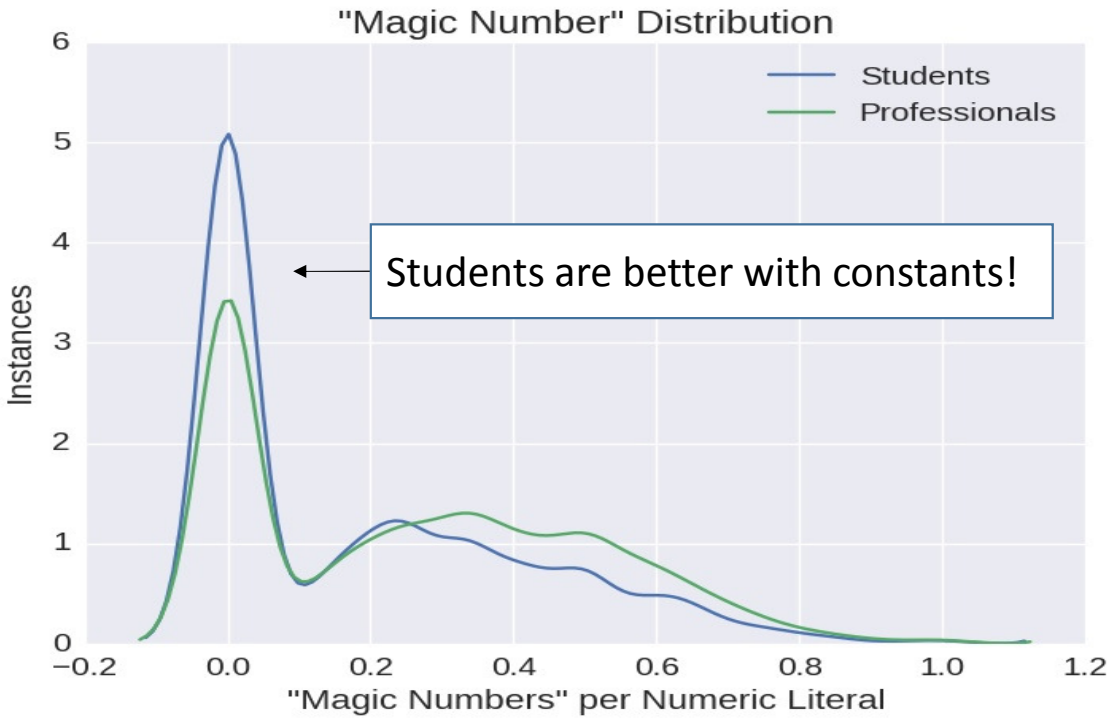
    `x` vs. `textLen`
  - **Limited redundancy**, pros write functions!

    `x+y*x-z` $=_p$ `a+b*a-c`
  - …



Data Story

Everybody needs `> ProCode( );`



NumPy : datetime.c

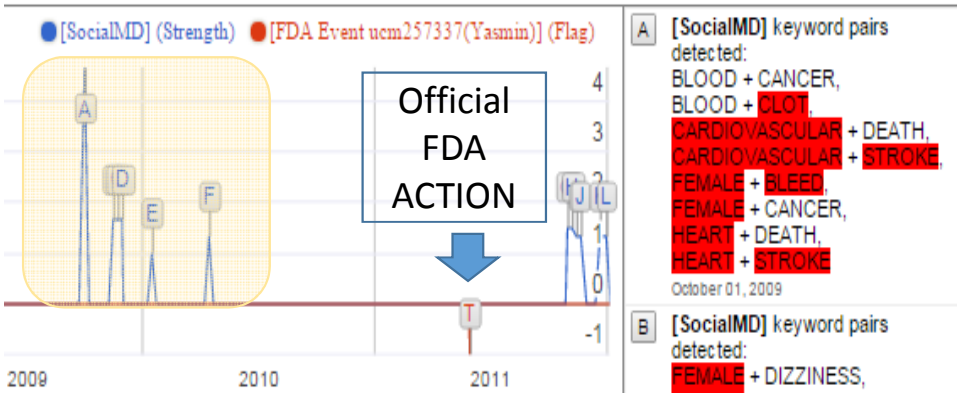
```
case NPY_FR_ms:
    ret = (((days * 24 +
            dts->hour) * 60 +
            dts->min) * 60 +
            dts->sec) * 1000 +
            dts->us / 1000;
    break;
```

Hard-Coded #s	Frequency
60	108
1000	45
1000000	38
24	36
2	29
…	…

Richard Beal

PhD Computer Science

Drug Adverse Reaction Detection



Parameterized Pattern Matching