

***Tibet -  
Table  
Structure  
Recognition  
Library***

<https://sourceforget.net/projects/tibet>

Haim Cohen



# Overview

- A library for table extraction
  - Written in C++
  - Regression tests
  - Open Source :  
<https://sourceforge.net/projects/tibet>
  - Treats the table as an “image” of characters
- 
-

# Simple Interface

```
-----  
Green   | 10 | 2.5  
Apple   |   |  
-----  
Banana  |   | 0.5  
-----  
Orange |1   |1.25  
-----
```

```
Tibet::Table t ;
```

```
std::cin >> t ;
```

```
std::size_t number_of_rows = t.rows() ; // 3
```

```
std::size_t number_of_cols = t.cols() ; // 3
```

```
std::string cell_contents = t.at( 2, 0 ) ; // Orange
```

# *Limitations*

- Only non-hierarchical tables.
  - No special cells (header) recognition.
  - Each input stream should have at most one table.
- 
-

# Input

```

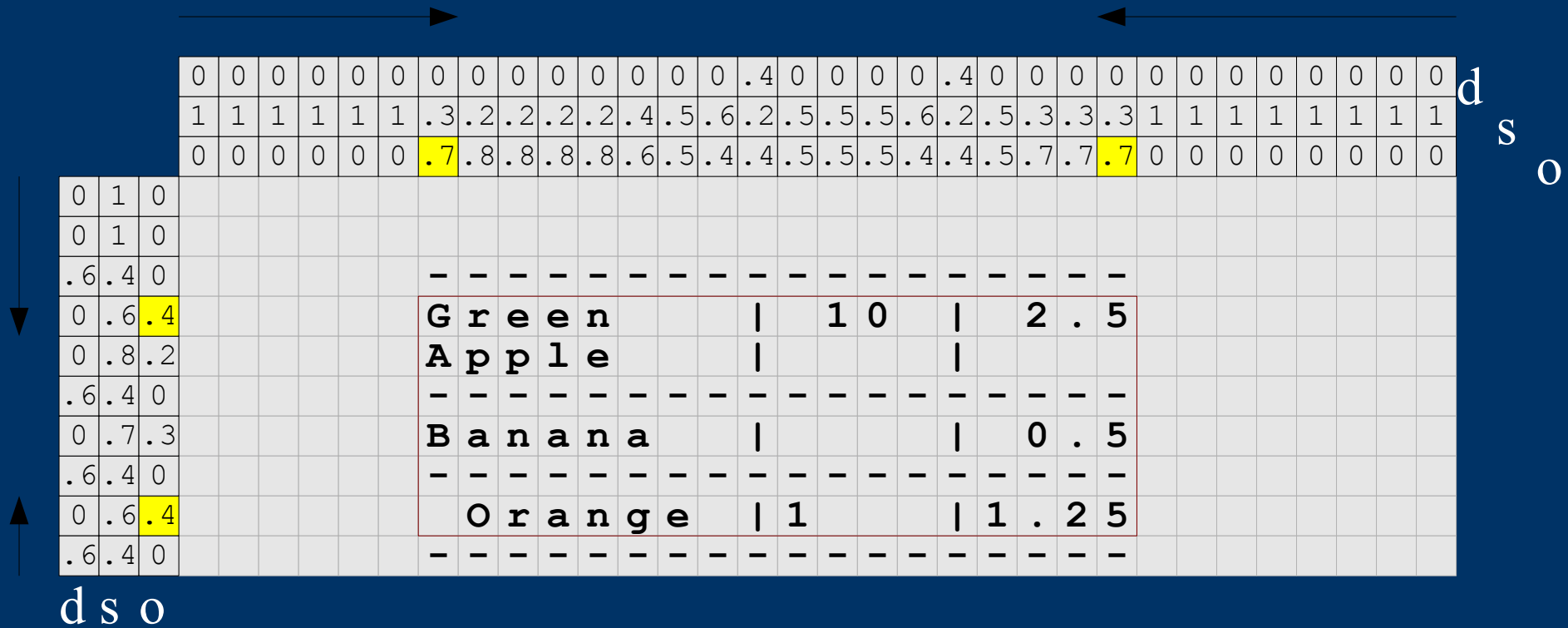
- - - - -
G r e e n   | 1 0 | 2 . 5
A p p l e   |   |
- - - - -
B a n a n a |   | 0 . 5
- - - - -
O r a n g e | 1 | 1 . 2 5
- - - - -

```



# Table Boundary Detection

other  $\geq 0.15$



# Second Histogram Calculation

|   |    |    |          |          |          |          |          |          |          |    |    |          |          |          |    |    |          |          |          |          |   |   |   |
|---|----|----|----------|----------|----------|----------|----------|----------|----------|----|----|----------|----------|----------|----|----|----------|----------|----------|----------|---|---|---|
|   |    |    | 0        | 0        | 0        | 0        | 0        | 0        | 0        | 0  | 0  | 0        | 0        | .6       | 0  | 0  | 0        | 0        | .6       | 0        | 0 | 0 | 0 |
|   |    |    | .2       | .2       | .2       | .2       | .2       | .4       | .5       | .6 | 0  | .5       | .5       | .5       | .6 | 0  | .5       | .2       | .2       | .2       |   |   |   |
|   |    |    | .8       | .8       | .8       | .8       | .8       | .6       | .5       | .4 | .4 | .5       | .5       | .5       | .4 | .4 | .5       | .8       | .8       | .8       |   |   |   |
| 0 | .3 | .7 | <b>G</b> | <b>r</b> | <b>e</b> | <b>e</b> | <b>n</b> |          |          |    |    |          | <b>1</b> | <b>0</b> |    |    |          | <b>2</b> | <b>.</b> | <b>5</b> |   |   |   |
| 0 | .6 | .4 | <b>A</b> | <b>p</b> | <b>p</b> | <b>l</b> | <b>e</b> |          |          |    |    |          |          |          |    |    |          |          |          |          |   |   |   |
| 1 | 0  | 0  | -        | -        | -        | -        | -        | -        | -        | -  | -  | -        | -        | -        | -  | -  | -        | -        | -        | -        | - | - | - |
| 0 | .7 | .3 | <b>B</b> | <b>a</b> | <b>n</b> | <b>a</b> | <b>n</b> | <b>a</b> |          |    |    |          |          |          |    |    |          | <b>0</b> | <b>.</b> | <b>5</b> |   |   |   |
| 1 | 0  | 0  | -        | -        | -        | -        | -        | -        | -        | -  | -  | -        | -        | -        | -  | -  | -        | -        | -        | -        | - | - | - |
| 0 | .3 | .7 |          | <b>O</b> | <b>r</b> | <b>a</b> | <b>n</b> | <b>g</b> | <b>e</b> |    |    | <b>1</b> |          |          |    |    | <b>1</b> | <b>.</b> | <b>2</b> | <b>5</b> |   |   |   |

d  
s  
o

d  
s  
o



# Table Frames Detection

delimiters  $\geq 0.3$

|   |    |    |          |          |          |          |          |          |          |    |    |          |          |          |    |    |          |          |          |          |
|---|----|----|----------|----------|----------|----------|----------|----------|----------|----|----|----------|----------|----------|----|----|----------|----------|----------|----------|
|   |    |    | 0        | 0        | 0        | 0        | 0        | 0        | 0        | 0  | .6 | 0        | 0        | 0        | 0  | .6 | 0        | 0        | 0        | 0        |
|   |    |    | .2       | .2       | .2       | .2       | .2       | .4       | .5       | .6 | 0  | .5       | .5       | .5       | .6 | 0  | .5       | .2       | .2       | .2       |
|   |    |    | .8       | .8       | .8       | .8       | .8       | .6       | .5       | .4 | .4 | .5       | .5       | .5       | .4 | .4 | .5       | .8       | .8       | .8       |
| 0 | .3 | .7 | <b>G</b> | <b>r</b> | <b>e</b> | <b>e</b> | <b>n</b> |          |          |    |    |          | <b>1</b> | <b>0</b> |    |    |          | <b>2</b> | <b>.</b> | <b>5</b> |
| 0 | .6 | .4 | <b>A</b> | <b>p</b> | <b>p</b> | <b>l</b> | <b>e</b> |          |          |    |    |          |          |          |    |    |          |          |          |          |
| 1 | 0  | 0  | -        | -        | -        | -        | -        | -        | -        | -  | -  | -        | -        | -        | -  | -  | -        | -        | -        | -        |
| 0 | .7 | .3 | <b>B</b> | <b>a</b> | <b>n</b> | <b>a</b> | <b>n</b> | <b>a</b> |          |    |    |          |          |          |    |    |          | <b>0</b> | <b>.</b> | <b>5</b> |
| 1 | 0  | 0  | -        | -        | -        | -        | -        | -        | -        | -  | -  | -        | -        | -        | -  | -  | -        | -        | -        | -        |
| 0 | .3 | .7 |          | <b>O</b> | <b>r</b> | <b>a</b> | <b>n</b> | <b>g</b> | <b>e</b> |    |    | <b>1</b> |          |          |    |    | <b>1</b> | <b>.</b> | <b>2</b> | <b>5</b> |

d

s

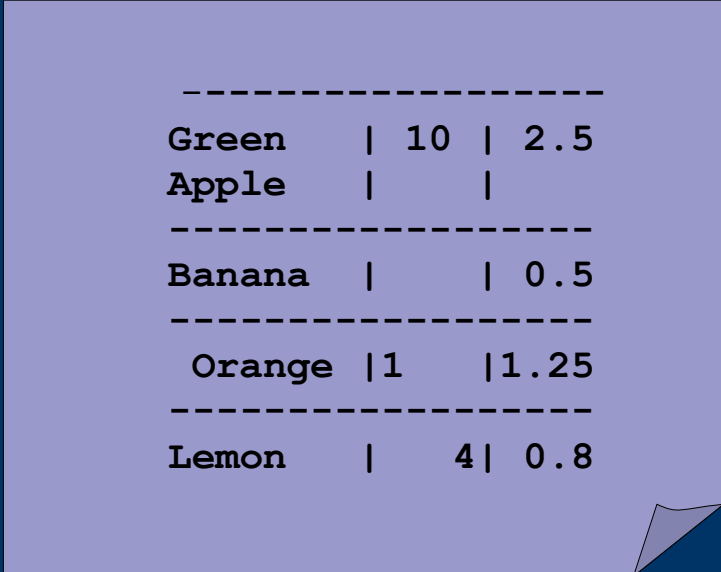
o

d s o



# Cells Extraction

```
Green   Apple | 10 | 2.5
Banana  |  | 0.5
Orange | 1 | 1.25
```



|        |    |      |
|--------|----|------|
| Green  | 10 | 2.5  |
| Apple  |    |      |
| Banana |    | 0.5  |
| Orange | 1  | 1.25 |
| Lemon  | 4  | 0.8  |

# *Evaluation*

- Complex tables ([www.fedstats.gov](http://www.fedstats.gov))
- No labeled tables
- Created a labeled data set
- ~ 90 %



# Real Life Data

## Missouri Soybean Production and Yield by District

| Area  | Acres Planted |       | Acres Harvested |       | Yield     |      | Production     |         |
|-------|---------------|-------|-----------------|-------|-----------|------|----------------|---------|
|       | 2004          | 2005  | 2004            | 2005  | 2004      | 2005 | 2004           | 2005    |
|       | -thousands-   |       | -thousands-     |       | -bu/acre- |      | -thousand bu.- |         |
|       |               |       |                 |       |           |      |                |         |
| NW    | 946.4         | 959   | 938.8           | 950   | 47.4      | 37   | 44,466         | 35,600  |
| NC    | 692.6         | 714   | 684.1           | 707   | 45.1      | 31   | 30,842         | 22,200  |
| NE    | 812.7         | 821   | 806.9           | 812   | 47.9      | 26   | 38,690         | 21,400  |
| WC    | 544.5         | 561   | 540.4           | 556   | 44.0      | 30   | 23,797         | 16,700  |
| C     | 531.7         | 535   | 528.2           | 530   | 47.1      | 24   | 24,887         | 12,800  |
| EC    | 338.0         | 347   | 335.6           | 344   | 45.6      | 24   | 15,307         | 8,400   |
| SW    | 156.0         | 158   | 154.9           | 156   | 34.4      | 24   | 5,329          | 3,800   |
| SC    | 35.9          | 41    | 35.5            | 40    | 37.5      | 34   | 1,331          | 1,350   |
| SE    | 942.2         | 964   | 935.6           | 955   | 41.2      | 36   | 38,551         | 34,300  |
| STATE | 5,000.0       | 5,100 | 4,960.0         | 5,050 | 45.0      | 31   | 223,200        | 156,550 |

# *What Next ?*

- Detect implicit frames
- Use spectral properties of the characters
- Table header detection
- Hierarchical tables



# Test Data Samples

```
|a|b|c|
-----
|1|2|3|
|-|-|-|
|z|z|z|
```

```
+ - - - + - - - +
|  A    |  B    |
+ - - - + - - - +
|  C    |  D    |
+ - - - + - - - +

|a|b|c|
=====
|1|2|3|
-----
|z|z|z|
-----
|x|y|z|
|!|!|!|
-----

-----
|1|2|3|4|
-----
|A|B|C|D|
-----
|@|%|^|!|
```

```
1 | 2
  |
-----
4 | 3
  |
```

```
+-----+-----+-----+-----+-----+
|      | 10    | 100   | 1000  | 10000 | all    |
+-----+-----+-----+-----+-----+
| 20   | 85.50  | 97.50 | 97.50 | 97.50 | 97.50  |
-----
| 100  | 88.00  | 96.50 | 100.00 | 100.00 | 100.00 |
-----
| 300  | 88.67  | 95.67 | 99.67  | 99.83  | 99.83  |
-----
| 500  | 87.20  | 96.20 | 99.70  | 100.00 | 99.70  |
+-----+-----+-----+-----+-----+
```