
# CUSTODES: Automatic Spreadsheet Cell Clustering and Smell Detection using Strong and Weak Features




**Shing-Chi Cheung** **Wanjun Chen** **Yepang Liu** **Chang Xu**

*The Hong Kong University of Science and Technology*

*Nanjing University*

# Spreadsheets are mission critical!
# Spreadsheets are popular (over 1 billion of users)

"Most business today rely on spreadsheets in some way. The multi-celled document is used heavily for finance and account, as well as supply chain, customer relationship and sales functions." … www.cio.com


Photograph by Gregor Schuster

# Spreadsheet mistakes are also popular!!

News › Politics › George Osborne

## The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

Charles Arthur and Phillip Inman
The Guardian, Thursday 18 April 2013 21.10 BST


## 88% of spreadsheets have errors

By Jeremy Olshan
Published: Apr 20, 2013 7:39 a.m. ET

2,308 1,299 137

F OPINION

Capital Flows
Contributor
*Guest commentary curated by Forbes*

OPINION 4/18/2013 @ 5:58AM | 11,891 views

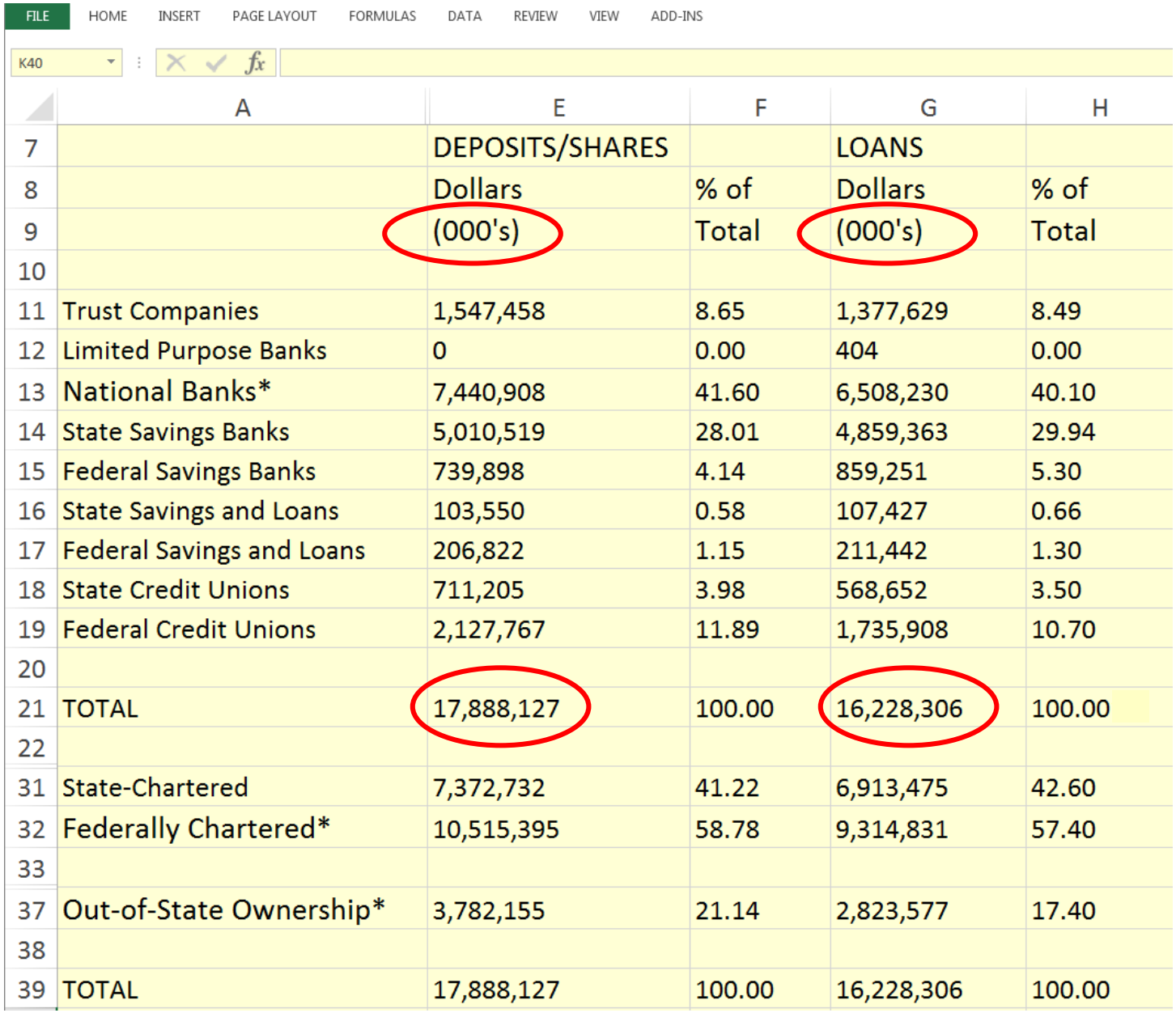
## That Reinhart and Rogoff Committed a Spreadsheet Error Completely Misses the Point

+ Comment Now + Follow Comments

An example of a **mission critical** spreadsheet.

- may concern an executive decision on **17 billion** dollars!

- extracted from the EUSES corpus



FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS

K40

| | A | E | F | G | H |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 20 | | | | | |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 22 | | | | | |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 38 | | | | | |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

Excel finds **no warnings** at these cells!

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS

K40

| | A | E | F | G | H |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 20 | | | | | |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 22 | | | | | |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 38 | | | | | |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

- Any hidden **problems**?
- Safe to **update** the values?
- Safe to **copy-and-paste** and make changes?

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS

K40

| | A | E | F | G | H |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 20 | | | | | |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 22 | | | | | |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 38 | | | | | |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

Can we detect problems in the **absence of functional specification**?

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS

K40

| | A | E | F | G | H |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 20 | | | | | |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 22 | | | | | |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 38 | | | | | |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

# Inspect spreadsheet in R1C1 reference format

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |


# Cluster **similarly computed cells** and identify **issues**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

# Cluster more similarly computed cells and identify more issues

**value given by formula =(RC[-1]/R21C[-1])*100**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 → 1.16 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

# Cluster more similarly computed cells and identify more issues

**values given by formula =(RC[-1]/R21C[-1])*100**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 → 1.16 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

# Cluster more similarly computed cells and identify more issues

**values given by formula =(RC[-1]/R21C[-1])*100**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 → 5.29 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 → 1.16 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 → 3.50 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

# **Not all** issues (smells) indicate **errors** in calculation

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 → 5.29 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 → 1.16 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 → 3.50 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*100 | 2225 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**Three SMELLS**

# Find more clusters

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 → 5.29 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 → 1.16 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 → 3.50 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**Should be similarly computed**

# A possible scenario of introducing smells

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 → 5.29 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | =(RC[-1]/R21C[-1])*100 → 1.15 | | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 → 3.50 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) → 100% | ]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

1.16

99.99%

**Should be similarly computed**

# Cluster more similarly computed cells

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**Should these four cells be also similarly computed?**

# The formulas have **different summation ranges**!!

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**=SUM(R[-10]C:R[-2]C)**

# The formulas have **different summation ranges**!!

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**=SUM(R[-10]C:R[-1]C)** **=SUM(R[-10]C:R[-2]C)**

# Table cannot be safely updated or copy-and-paste

**Implicitly assume this row must be empty OR**

**No more data rows will be added here**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | SECURITIES/LOANS | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**=SUM(R[-10]C:R[-1]C)** **=SUM(R[-10]C:R[-2]C)**

# No immediate calculation errors **BUT …**

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | | | | |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | | | | |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**Spreadsheet smells can induce problems to future spreadsheet maintenance and reuse like code smells**

# Challenges: Do contiguous cells always form a cluster?

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

# Challenges: Do contiguous cells always form a cluster?

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2323577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**Should they form a cluster?**

# Challenges: Do contiguous cells always form a cluster?

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C) | | R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

adopt the formula here?

# Challenges: Do contiguous cells always form a cluster?

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | | | | |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | | | | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**References non-numeric cells for computation!**

**=(RC[-1]/R[2]C[-1])*R[-16]C)**

# Challenges: Do contiguous cells always form a cluster?

| | | | | | |
|---|---|---|---|---|---|
| 7 | | DEPOSITS/SHARES | | LOANS | |
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**form a cluster**

# Challenges of clustering

**Observations**:

- Cell clusters can be **discontiguous**
- Cells in a cluster may reside in **different rows** or **columns**
- Users often have **different tabulation styles**

| | | | | |
|---|---|---|---|---|
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**form a cluster**

# Challenges of clustering

**Observations**:

- Cell clusters can be **discontiguous**
- Cells in a cluster may reside in **different rows** or **columns**
- Users often have **different tabulation styles**

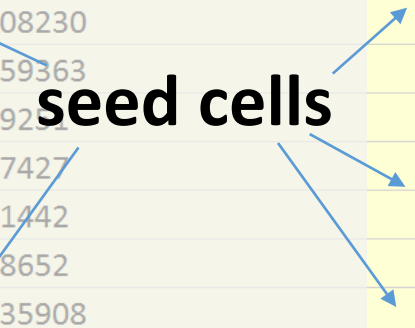
**Can we apply clustering techniques to solve the problem?**
**However, different clusters have different features …**

| | | | | |
|---|---|---|---|---|
| 21 | TOTAL | =SUM(R[-10]C:R[-1]C) | 100 | =SUM(R[-10]C:R[-1]C) | =SUM(R[-10]C:R[-2]C) |
| 22 | | | | | |
| 31 | State-Chartered | =SUM(R[-20]C,R[-19]C,R[-17]C, | =SUM(R[-20]C,R[-19]C,R[-17]C | =SUM(R[-20]C,R[-19]C,R[-17]C,R | =SUM(R[-20]C,R[-19]C,R[-17]C,R |
| 32 | Federally Chartered* | =SUM(R[-19]C,R[-17]C,R[-15]C, | =SUM(R[-19]C,R[-17]C,R[-15]C | =SUM(R[-19]C,R[-17]C,R[-15]C,R | =R[-19]C+R[-17]C+R[-15]C+R[-13 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3782155 | =(RC[-1]/R[2]C[-1])*R[-16]C | 2823577 | =(RC[-1]/R[2]C[-1])*100 |
| 38 | | | | | |
| 39 | TOTAL | =R[-5]C | =R[-5]C | =R[-5]C | =R[-5]C |

**form a cluster**

# Two-stage clustering: stage 1 (strong features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |




- Formula cells in the **same cluster** perform the **same computation**
- Model various **formula similarities** as **strong features**
- Identify cells with very similar formulas as **seed cells**

# Two-stage clustering: stage 1 (strong features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508230 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 882251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**a seed cluster**

- Formula cells in the **same cluster** perform the **same computation**
- Model various **formula similarities** as **strong features**
- Identify cells with very similar formulas as **seed cells**
- Group similar seed cells into a seed cluster

# Two-stage clustering: stage 2 (weak features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 65082 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211442 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**labels**

**What are their weak features?**

- **Weak features** are characteristics **specific to a (seed) cluster**
- Examples are: **labels**, **layout style**, row/column **indices**, horizontal/vertical **gaps** between cells, and so on
- Extract **applicable weak features** from a given **seed cluster**

# Two-stage clustering: stage 2 (weak features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739808 | =(RC[-1]/R21C[-1])*100 | 859 | 5.3 |
| 16 | State Savings and Loans | | =(RC[-1]/R21C[-1])*100 | 10742 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2137377 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**Likely have the same column label "% of and Total"**

**What are their weak features?**

- **Weak features** are characteristics **specific to a (seed) cluster**
- Examples are: **labels**, **layout style**, **data value indicators**, horizontal/vertical **gaps** between cells, and so on
- Extract **applicable weak features** from a given **seed cluster**

**Likely reside in two specific columns**
**Likely reside between row 11 and 19**

# Two-stage clustering: stage 2 (weak features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | [illegible] | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508[illegible] | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | [illegible] | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739808 | =(RC[-1]/R21C[-1])*100 | 859[illegible] | 5.3 |
| 16 | State Savings and Loans | [illegible] | =(RC[-1]/R21C[-1])*100 | 10742[illegible] | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 11[illegible] | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2137[illegible] | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**Likely have the same column label "% of and Total"**

**What are their weak features?**

**Likely reside in two specific columns**
**Likely reside between row 11 and 19**

- **Weak features** are characteristics **specific to a cluster**
- Examples are: **labels**, **layout style**, **cell value indicators**, horizontal/vertical **gaps** between cells, and so on
- Extract **applicable weak features** from a given **seed cluster**
- Cluster **more** cells **sharing** the same extracted **weak features**

# Two-stage clustering: stage 2 (weak features)

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 173 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**share the same weak features with the seed cells**

- Adapt a **bootstrapping algorithm** proposed for computational linguistics [Pantel & Pennacchiotti, ACL 2006]
- Please refer to the paper for the **details of cell similarity functions**, **feature modeling** and **clustering algorithm**

# A way to use clustering results: smell detection

| 7 | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508238 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211440 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**Detect smells as outliers**

- Determine the **type of smells** to be detected, e.g., computational smells
- Model the **feature space** used for outlier detection: e.g., R1C1 expressions, AST structures, cell dependency trees, and so on
- **Detect outliers**, e.g., a Local Outlier Factor (LOF) value > 1.0

# A way to use clustering results: smell detection

| 7 | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1547458 | =(RC[-1]/R21C[-1])*100 | 1377629 | =(RC[-1]/R21C[-1])*100 |
| 12 | Limited Purpose Banks | 0 | =(RC[-1]/R21C[-1])*100 | 404 | =(RC[-1]/R21C[-1])*100 |
| 13 | National Banks* | 7440908 | =(RC[-1]/R21C[-1])*100 | 6508238 | =(RC[-1]/R21C[-1])*100 |
| 14 | State Savings Banks | 5010519 | =(RC[-1]/R21C[-1])*100 | 4859363 | =(RC[-1]/R21C[-1])*100 |
| 15 | Federal Savings Banks | 739898 | =(RC[-1]/R21C[-1])*100 | 859251 | 5.3 |
| 16 | State Savings and Loans | 103550 | =(RC[-1]/R21C[-1])*100 | 107427 | =(RC[-1]/R21C[-1])*100 |
| 17 | Federal Savings and Loans | 206822 | 1.15 | 211440 | =(RC[-1]/R21C[-1])*100 |
| 18 | State Credit Unions | 711205 | =(RC[-1]/R21C[-1])*100 | 568652 | 3.5 |
| 19 | Federal Credit Unions | 2127767 | =(RC[-1]/R21C[-1])*100 | 1735908 | =(RC[-1]/R21C[-1])*100 |
| 20 | | | | | |

**Detect smells as outliers**

Possible variations:

- Increase precision by **combining** outlier detection results using **multiple feature spaces**
- Combine semi-supervised training or other learning techniques

# CUSTODES: Implementation over Apache POI

- Mark clusters in different colors
- Mark smells as comments with formula suggestion

Download link:
http://sccpu2.cse.ust.hk/custodes/

=(RC[-1]/R21C[-1])*100

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| | | Dollars (000's) | % of Total | Dollars (000's) | % of Total |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

# Evaluate CUSTODES on 70 randomly sampled EUSES Spreadsheets

| CATEGORY | # WORKSHEETS | # CELLS | # FORMULA CELLS | # CLUSTERS | # SMELLY CELLS |
|---|---|---|---|---|---|
| CS101 | 1 | 106 | 40 | 8 | 3 |
| DATABASE | 60 | 42,688 | 6,973 | 547 | 1,206 |
| FINANCIAL | 102 | 54,734 | 5,692 | 533 | 477 |
| FORMS3 | 5 | 1,774 | 734 | 35 | 12 |
| GRADES | 30 | 23,998 | 2,571 | 73 | 124 |
| HOMEWORK | 23 | 12,137 | 3,878 | 150 | 50 |
| INVENTORY | 35 | 17,082 | 1,927 | 125 | 59 |
| MODELING | 35 | 36,508 | 4,901 | 139 | 43 |
| TOTAL | 291 | 189,027 | 26,716 | 1,610 | 1,974 |

# Smell detection results compared with existing smell/error detection techniques

| Category | Smell Detection Results of Different Techniques | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CUSTODES | | AmCheck | | UCheck | | Dimension Inference | | Excel 2013's Error Checking | |
| | Detected | True | Detected | True | Detected | True | Detected | True | Detected | True |
| cs101 | 3 | 3 | 6 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| database | 1,116 | 1,066 | 823 | 790 | 158 | 0 | 863 | 4 | 563 | 18 |
| financial | 651 | 317 | 502 | 287 | 0 | 0 | 151 | 1 | 1,204 | 70 |
| forms3 | 29 | 10 | 73 | 1 | 0 | 0 | 0 | 0 | 464 | 2 |
| grades | 316 | 94 | 79 | 78 | 0 | 0 | 11 | 0 | 322 | 8 |
| homework | 95 | 47 | 109 | 16 | 0 | 0 | 370 | 0 | 1,238 | 6 |
| inventory | 144 | 27 | 372 | 37 | 34 | 0 | 190 | 2 | 391 | 17 |
| modeling | 89 | 19 | 199 | 21 | 12 | 1 | 112 | 4 | 798 | 22 |
| Total | 2,443 | 1,583 (71.68%) | 2,163 | 1,231 (59.51%) | 204 | 1 (0.09%) | 1,697 | 11 (0.60%) | 4,981 | 143 (4.11%) |

f-measure

with all checking rules enabled

# Opportunities and future work

=SUM(RC[-4]:RC[-1])

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 59 | Fiscal Year 2004 | QTR 1 | QTR 2 | QTR 3 | QTR 4 | YTD2004 |
| 62 | | | | | | |
| 63 | (2) GAAP Net Income | 7.5 | =+R[-51]C | =+R[-51]C | | =SUM(RC[-4]:RC[-1]) |
| 64 | Non-recurring tax benef | 0 | -19.7 | 0 | | =SUM(RC[-4]:RC[-1]) |
| 65 | Pro Forma Net Income | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |
| 66 | | | | | | |
| 67 | (3) GAAP Earnings Per Sh | 0.07 | =+R[-54]C ( 0.29 ) | =+R[-54]C ( 0.2 ) | | 0.55 |
| 68 | Non-recurring tax benef | 0 | -0.18 | 0 | | -0.17 |
| 69 | Pro Forma Earnings Pe | =SUM(R[-2]C:R[-2]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |

successfully cluster these cells and suggest a fixing formula

# Opportunities and future work

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 59 | Fiscal Year 2004 | QTR 1 | QTR 2 | QTR 3 | QTR 4 | YTD2004 |
| 62 | | | | | | |
| 63 | (2) GAAP Net Income | 7.5 | =+R[-51]C | =+R[-51]C | | =SUM(RC[-4]:RC[-1]) |
| 64 | Non-recurring tax benef | 0 | -19.7 | 0 | | =SUM(RC[-4]:RC[-1]) |
| 65 | Pro Forma Net Income | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |
| 66 | | | | | | |
| 67 | (3) GAAP Earnings Per Sh | 0.07 | =+R[-54]C ( 0.29 ) | =+R[-54]C ( 0.2 ) | | 0.55 |
| 68 | Non-recurring tax benef | 0 | -0.18 | 0 | | -0.17 |
| 69 | Pro Forma Earnings Pe | =SUM(R[-2]C:R[-2]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |

- More **accurate** clustering by comparing which clusters offer a formula closer to the existing value?
- **More features** and combine results of multiple feature spaces?
- **Cluster data** using formulas identified?

# Opportunities and future work

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 59 | Fiscal Year 2004 | QTR 1 | QTR 2 | QTR 3 | QTR 4 | YTD2004 |
| 62 | | | | | | |
| 63 | (2) GAAP Net Income | 7.5 | =+R[-51]C | =+R[-51]C | | =SUM(RC[-4]:RC[-1]) |
| 64 | Non-recurring tax benef | 0 | -19.7 | 0 | | =SUM(RC[-4]:RC[-1]) |
| 65 | Pro Forma Net Income | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |
| 66 | | | | | | |
| 67 | (3) GAAP Earnings Per Sh | 0.07 | =+R[-54]C ( 0.29 ) | =+R[-54]C ( 0.2 ) | | 0.55 |
| 68 | Non-recurring tax benef | 0 | -0.18 | 0 | | -0.17 |
| 69 | Pro Forma Earnings Pe | =SUM(R[-2]C:R[-2]C) | =SUM(R[-2]C:R[-1]C) | =SUM(R[-2]C:R[-1]C) | | =SUM(R[-2]C:R[-1]C) |

- **New applications** of clustering results?
- **Test prioritization** based on clustering fuzziness and/or outlier suspiciousness?
- Automatic **test oracle generation**?
- Automatic **test generation** to confirm maintenance issues?

# Invitation to our presentation at SEIP track this afternoon

**03:00 – 03:20 PM (*Room: Wedgwood*)**
**VEnron: A Versioned Spreadsheet Corpus and Related Evolution Analysis**

**Wensheng Dou, Liang Xu, Shing-Chi Cheung, Chushu Gao, Jun Wei and Tao Huang**

# Q&A

*Maintenance issue*

*Missing formula*

*Inconsistent calculation*

| | | DEPOSITS/SHARES | | LOANS | |
|---|---|---|---|---|---|
| 8 | | Dollars | % of | Dollars | % of |
| 9 | | (000's) | Total | (000's) | Total |
| 10 | | | | | |
| 11 | Trust Companies | 1,547,458 | 8.65 | 1,377,629 | 8.49 |
| 12 | Limited Purpose Banks | 0 | 0.00 | 404 | 0.00 |
| 13 | National Banks* | 7,440,908 | 41.60 | 6,508,230 | 40.10 |
| 14 | State Savings Banks | 5,010,519 | 28.01 | 4,859,363 | 29.94 |
| 15 | Federal Savings Banks | 739,898 | 4.14 | 859,251 | 5.30 |
| 16 | State Savings and Loans | 103,550 | 0.58 | 107,427 | 0.66 |
| 17 | Federal Savings and Loans | 206,822 | 1.15 | 211,442 | 1.30 |
| 18 | State Credit Unions | 711,205 | 3.98 | 568,652 | 3.50 |
| 19 | Federal Credit Unions | 2,127,767 | 11.89 | 1,735,908 | 10.70 |
| 20 | | | | | |
| 21 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |
| 22 | | | | | |
| 31 | State-Chartered | 7,372,732 | 41.22 | 6,913,475 | 42.60 |
| 32 | Federally Chartered* | 10,515,395 | 58.78 | 9,314,831 | 57.40 |
| 33 | | | | | |
| 37 | Out-of-State Ownership* | 3,782,155 | 21.14 | 2,823,577 | 17.40 |
| 38 | | | | | |
| 39 | TOTAL | 17,888,127 | 100.00 | 16,228,306 | 100.00 |

URL: http://sccpu2.cse.ust.hk/custodes/

Email: scc@cse.ust.hk