Clustering header categories extracted from web tables*

	First author	Second author	Third author	Fourth author
Author	George Nagy	David W. Embley	Mukkai Krishnamoorthy+	Sharad Seth
Affiliation	RPI	BYU	RPI	UNL
Email	nagy@ecse.rpi.edu	embley@cs.byu.edu	moorthy@cs.rpi.edu	sethcse@unl.edu
Role	senior programmer	relations	RDF and Protégé	algebra & algorithms

* SPIE DR&R February 12, 2015, We gratefully acknowledge the DR&R reviewers' suggestions.

+ Mukkai acknowledges the help of Dr. Ravi Palla with Protégé.

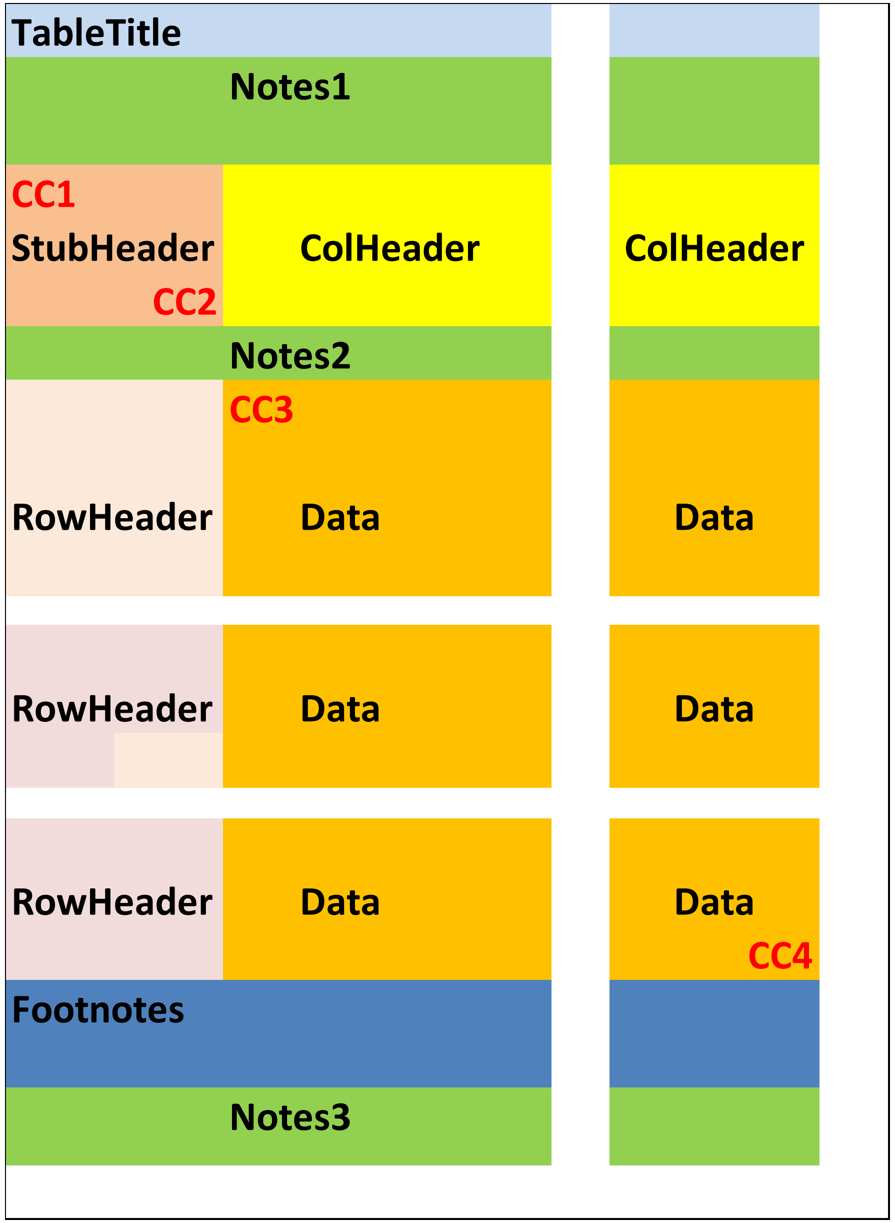
Agenda

- Prior work
- Well-formed tables
- Algorithmic category header extraction
- Clustering scheme
- Clustering results

Prior Work

- Clustering (numerical taxonomy, vector/graph, flat/hierarchical, fuzzy/crisp, SVM)
- Categories (data cube view)
- Physical structure extraction (mostly on rules/unruled scanned and ASCII tables)
- Logical structure extraction (HTML tables and spreadsheets)
    - Structure description trees, conditional random fields, grammars, syntactic coherency, web of concepts, Hurst, Pyvk, e Silva, Yahoo, CiteSeer) , BIG DATA,
    - Yahoo, CiteSeer, Google, Cafarella, Samet at al. row-by-row analysis
    - Current work mostly outside the DR&R community
- TANGO
    - Surveys, Egregious tables, Segmentation via MIPS, Factoring out categories, Lists, Interactive GT (DAS 2010, EIA 2011, DR&R 2012, ICPR 2012, ICDAR 2013, GREC 2013, DAS 2014, ICPR 2014)

Layout of a Well-Formed Table (WFT)


TableTitle

Notes1

CC1

StubHeader

CC2

ColHeader

ColHeader

Notes2

CC3

RowHeader

Data

Data

RowHeader

Data

Data

RowHeader

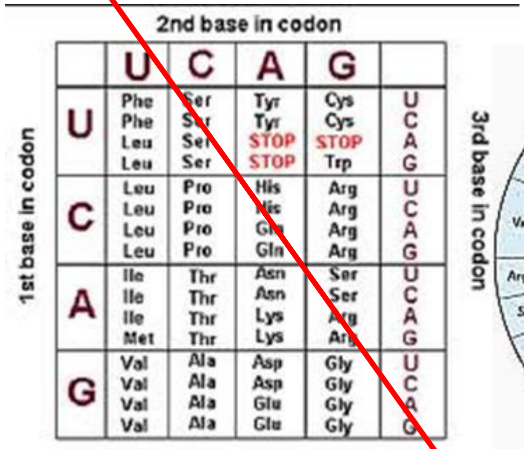
Data

Data

CC4

Footnotes

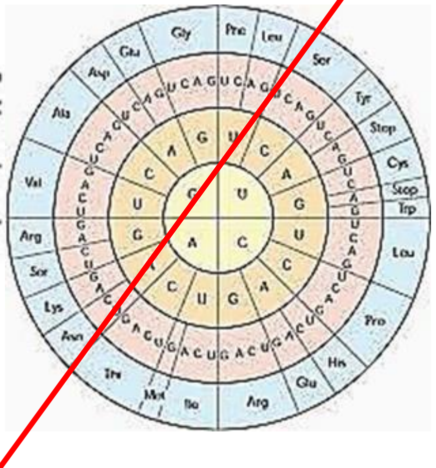
Notes3

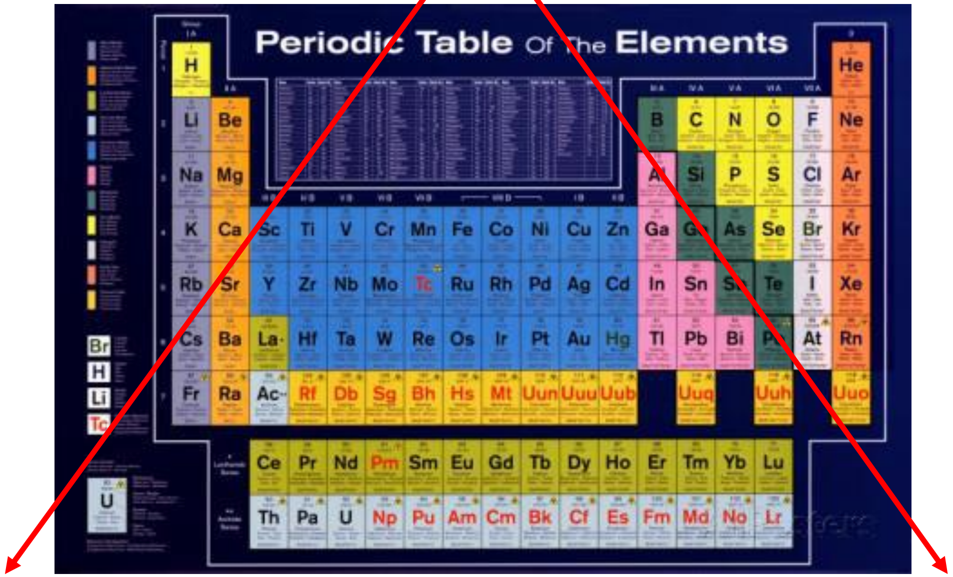

2nd base in codon

1st base in codon

3rd base in codon

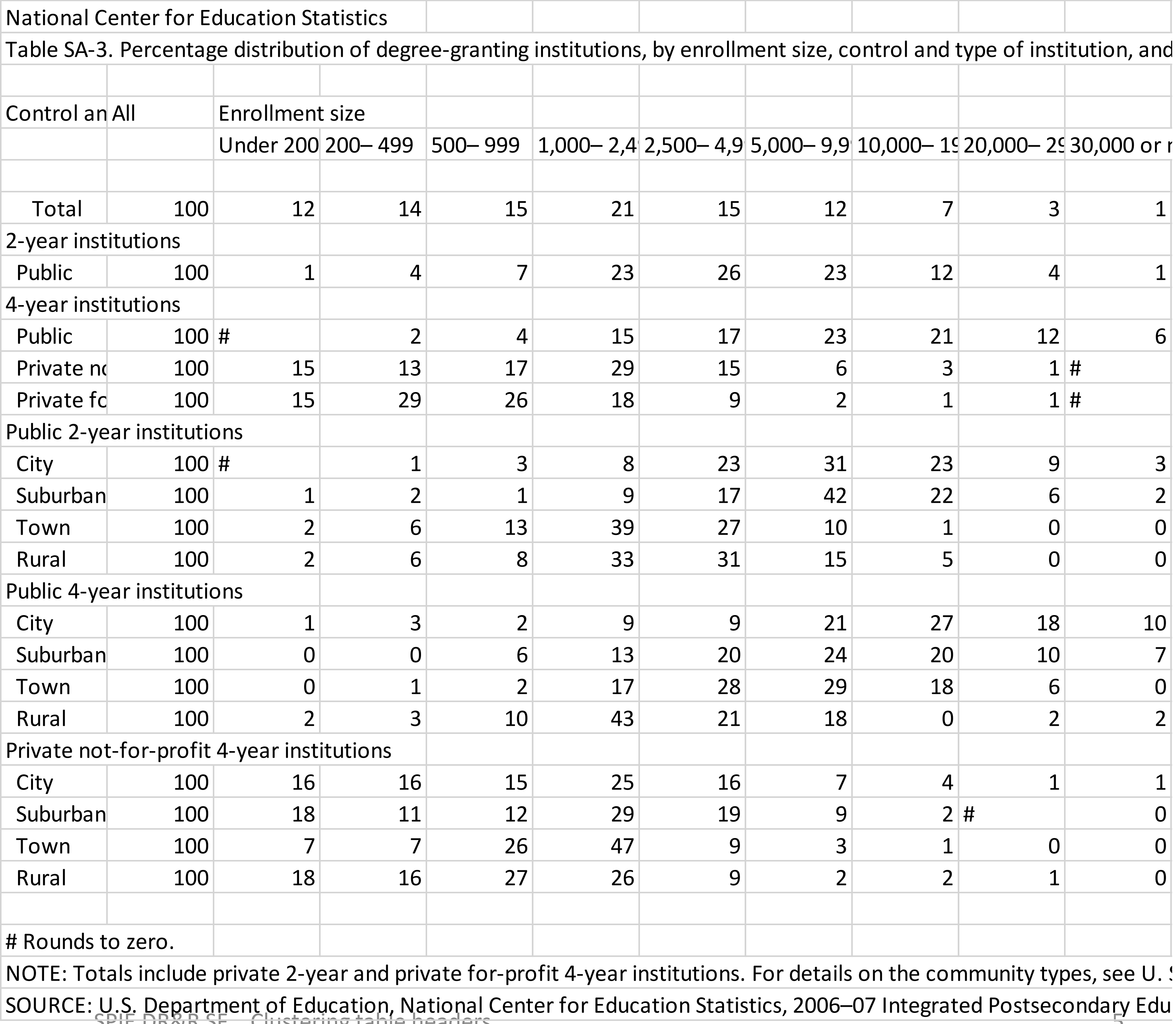
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G




Periodic Table of the Elements

H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt	Uun	Uuu	Uub		Uuq		Uuh		Uuo
		Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu		
		Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr		

Example WFT #1



National Center for Education Statistics										
Table SA-3. Percentage distribution of degree-granting institutions, by enrollment size, control and type of institution, and										
Control an	All	Enrollment size								
		Under 200	200– 499	500– 999	1,000– 2,4	2,500– 4,9	5,000– 9,9	10,000– 1	20,000– 2	30,000 or r
Total	100	12	14	15	21	15	12	7	3	1
2-year institutions										
Public	100	1	4	7	23	26	23	12	4	1
4-year institutions										
Public	100	#	2	4	15	17	23	21	12	6
Private nc	100	15	13	17	29	15	6	3	1	#
Private fc	100	15	29	26	18	9	2	1	1	#
Public 2-year institutions										
City	100	#	1	3	8	23	31	23	9	3
Suburban	100	1	2	1	9	17	42	22	6	2
Town	100	2	6	13	39	27	10	1	0	0
Rural	100	2	6	8	33	31	15	5	0	0
Public 4-year institutions										
City	100	1	3	2	9	9	21	27	18	10
Suburban	100	0	0	6	13	20	24	20	10	7
Town	100	0	1	2	17	28	29	18	6	0
Rural	100	2	3	10	43	21	18	0	2	2
Private not-for-profit 4-year institutions										
City	100	16	16	15	25	16	7	4	1	1
Suburban	100	18	11	12	29	19	9	2	#	0
Town	100	7	7	26	47	9	3	1	0	0
Rural	100	18	16	27	26	9	2	2	1	0
# Rounds to zero.										
NOTE: Totals include private 2-year and private for-profit 4-year institutions. For details on the community types, see U. S										
SOURCE: U.S. Department of Education, National Center for Education Statistics, 2006–07 Integrated Postsecondary Edu										

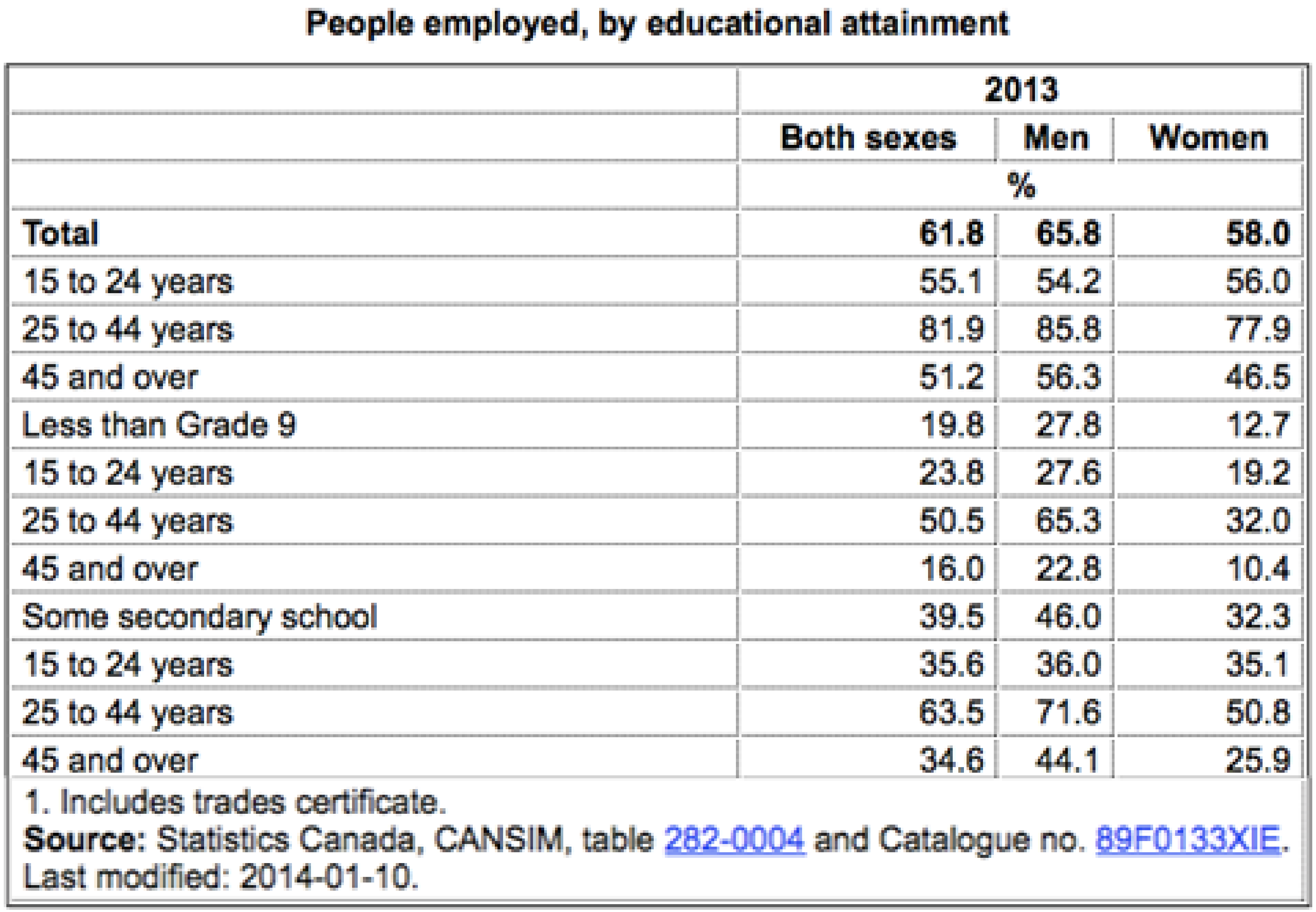
Example WFT #2 (two-category row header)

Both sexes, Men, Women

Total
Less than Grade 9
Some secondary school

×

15 to 24 years
25 to 44 years
45 and over



People employed, by educational attainment

	2013		
	Both sexes	Men	Women
	%		
Total	**61.8**	**65.8**	**58.0**
15 to 24 years	55.1	54.2	56.0
25 to 44 years	81.9	85.8	77.9
45 and over	51.2	56.3	46.5
Less than Grade 9	19.8	27.8	12.7
15 to 24 years	23.8	27.6	19.2
25 to 44 years	50.5	65.3	32.0
45 and over	16.0	22.8	10.4
Some secondary school	39.5	46.0	32.3
15 to 24 years	35.6	36.0	35.1
25 to 44 years	63.5	71.6	50.8
45 and over	34.6	44.1	25.9

1. Includes trades certificate.
Source: Statistics Canada, CANSIM, table 282-0004 and Catalogue no. 89F0133XIE.
Last modified: 2014-01-10.

Example ~WFT #3 (“crooked” column header)

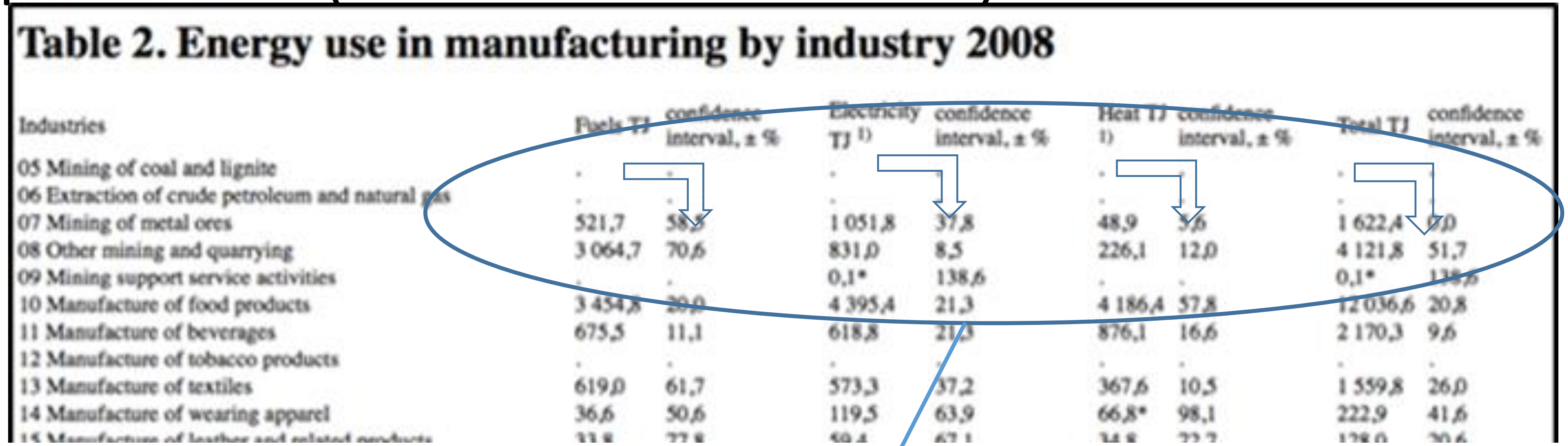


Table 2. Energy use in manufacturing by industry 2008

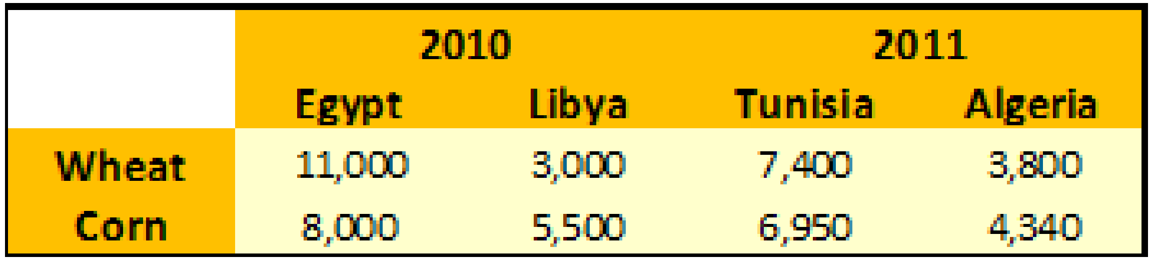
Industries	Fuels TJ	confidence interval, ± %	Electricity TJ [1)]	confidence interval, ± %	Heat TJ [1)]	confidence interval, ± %	Total TJ	confidence interval, ± %
05 Mining of coal and lignite	.	.	.	.	.	.	.	.
06 Extraction of crude petroleum and natural gas	.	.	.	.	.	.	.	.
07 Mining of metal ores	521,7	58,5	1 051,8	37,8	48,9	5,6	1 622,4	0,0
08 Other mining and quarrying	3 064,7	70,6	831,0	8,5	226,1	12,0	4 121,8	51,7
09 Mining support service activities	.	.	0,1*	138,6	.	.	0,1*	138,6
10 Manufacture of food products	3 454,8	20,0	4 395,4	21,3	4 186,4	57,8	12 036,6	20,8
11 Manufacture of beverages	675,5	11,1	618,8	21,3	876,1	16,6	2 170,3	9,6
12 Manufacture of tobacco products	.	.	.	.	.	.	.	.
13 Manufacture of textiles	619,0	61,7	573,3	37,2	367,6	10,5	1 559,8	26,0
14 Manufacture of wearing apparel	36,6	50,6	119,5	63,9	66,8*	98,1	222,9	41,6
15 Manufacture of leather and related products	33,8	77,8	59,4	67,1	34,8	72,7	128,0	20,6

Fuels TJ	confidence interval, ± %	Electricity TJ [1)]	confidence interval, ± %	Heat TJ[1)]	confidence interval, ± %	Total TJ	confidence interval, ± %
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
521,7	58,5	1 051,8	37,8	48,9	5,6	1 622,4	0,0
3 064,7	70,6	831,0	8,5	226,1	12,0	4 121,8	51,7

Four equivalent toy tables rendered before conversion to CSV format

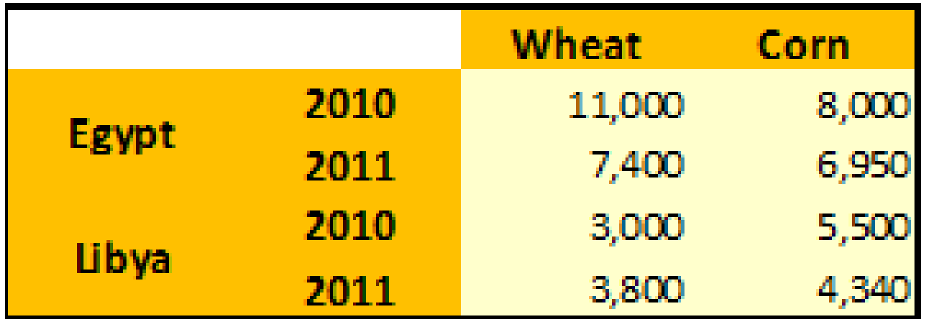
Note the spanning cells.

Table *12a* . Agricultural Production



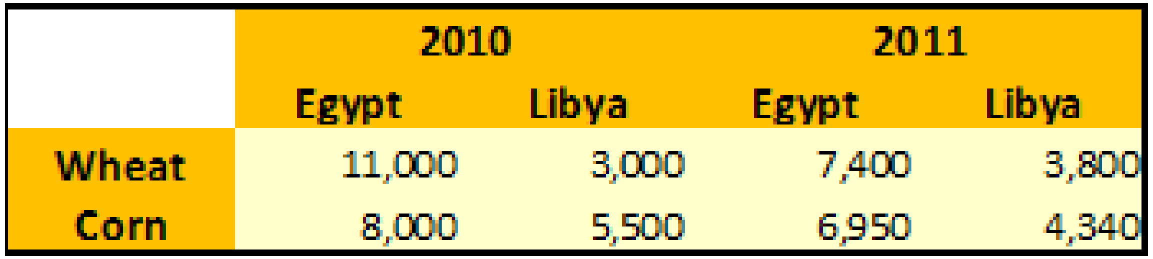
	2010		2011	
	Egypt	Libya	Tunisia	Algeria
Wheat	11,000	3,000	7,400	3,800
Corn	8,000	5,500	6,950	4,340

Table *12b* . Agricultural Production



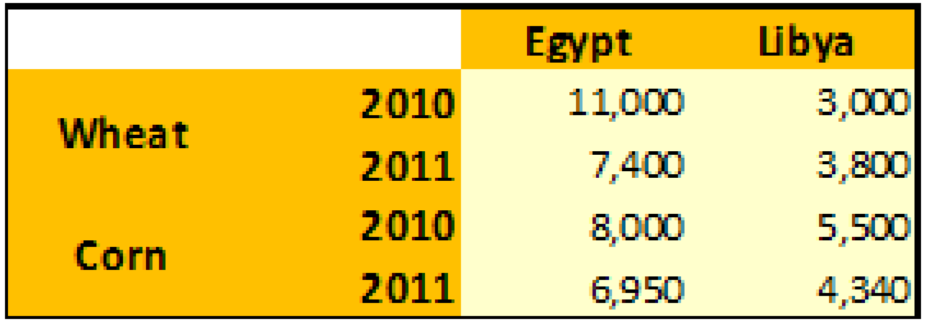
		Wheat	Corn
Egypt	2010	11,000	8,000
	2011	7,400	6,950
Libya	2010	3,000	5,500
	2011	3,800	4,340

Table *12c* . Agricultural Production



	2010		2011	
	Egypt	Libya	Egypt	Libya
Wheat	11,000	3,000	7,400	3,800
Corn	8,000	5,500	6,950	4,340

Table *12d.* Agricultural Production



		Egypt	Libya
Wheat	2010	11,000	3,000
	2011	7,400	3,800
Corn	2010	8,000	5,500
	2011	6,950	4,340

The same tables rendered after conversion to CSV format

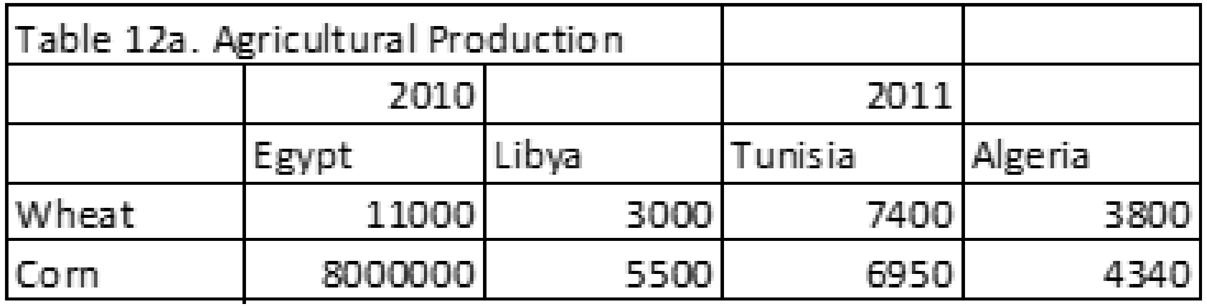


Table 12a. Agricultural Production				
	2010		2011	
	Egypt	Libya	Tunisia	Algeria
Wheat	11000	3000	7400	3800
Corn	8000000	5500	6950	4340

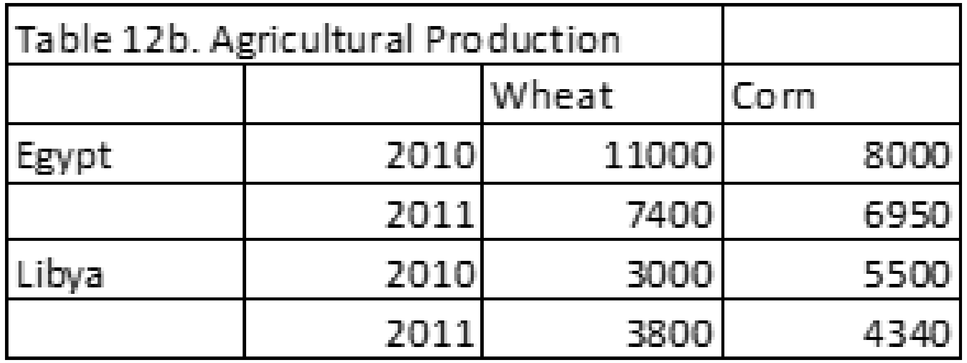


Table 12b. Agricultural Production			
		Wheat	Corn
Egypt	2010	11000	8000
	2011	7400	6950
Libya	2010	3000	5500
	2011	3800	4340

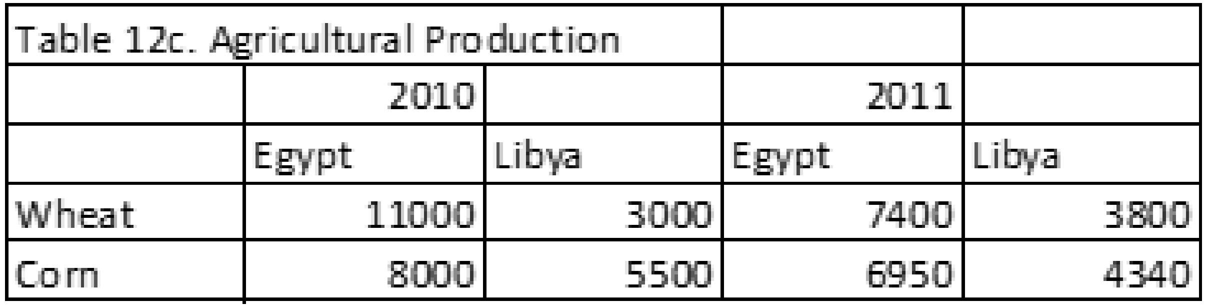


Table 12c. Agricultural Production				
	2010		2011	
	Egypt	Libya	Egypt	Libya
Wheat	11000	3000	7400	3800
Corn	8000	5500	6950	4340

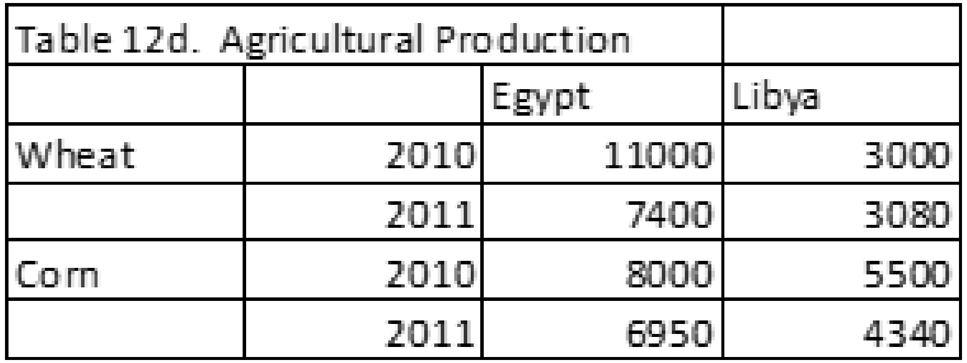


Table 12d. Agricultural Production			
		Egypt	Libya
Wheat	2010	11000	3000
	2011	7400	3080
Corn	2010	8000	5500
	2011	6950	4340

Notepad (.txt) display of Table 12d

Six lines, three commas per line

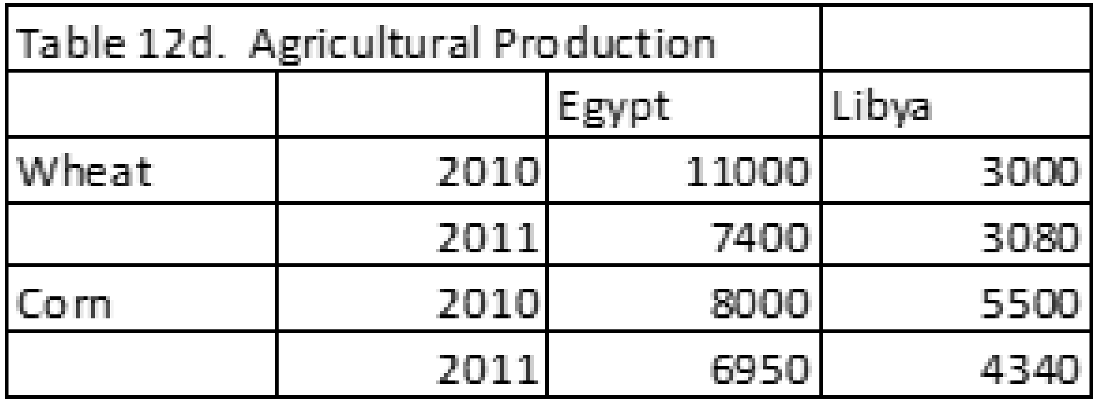


Table 12d. Agricultural Production			
		Egypt	Libya
Wheat	2010	11000	3000
	2011	7400	3080
Corn	2010	8000	5500
	2011	6950	4340

```
Table 12d.  Agricultural Production,,,
,,Egypt,Libya
Wheat,2010,11000,3000
,2011,7400,3080
Corn,2010,8000,5500
,2011,6950,4340
```

Table 12c after refilling split spanning-cell contents

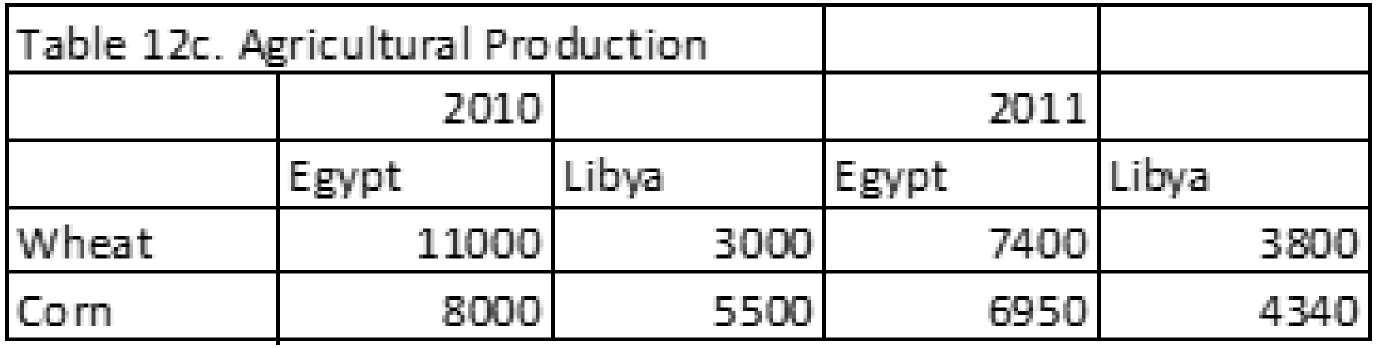


Table 12c. Agricultural Production				
	2010		2011	
	Egypt	Libya	Egypt	Libya
Wheat	11000	3000	7400	3800
Corn	8000	5500	6950	4340

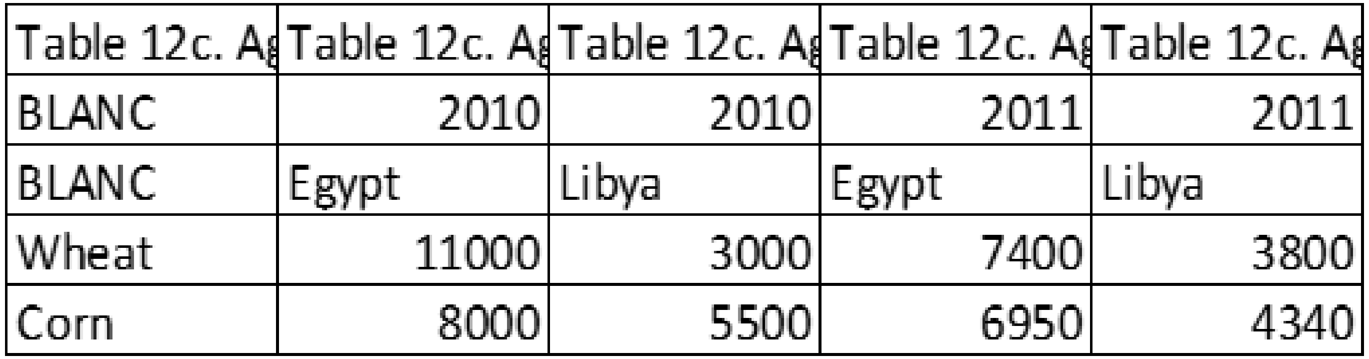


Table 12c. Ag	Table 12c. Ag	Table 12c. Ag	Table 12c. Ag	Table 12c. Ag
BLANC	2010	2010	2011	2011
BLANC	Egypt	Libya	Egypt	Libya
Wheat	11000	3000	7400	3800
Corn	8000	5500	6950	4340

Segmentation strategy !

MIPS (minimum index point search) algorithm finds smallest number of

rows for *unique* column-header paths

columns for *unique* row-header paths

(*ICDAR 2013*)

Partial output of segmentation program and corresponding GT

```
TableID     CC1    CC2   CC3    CC4    TableID      CC1    CC2   CC3    CC4
…
C10021.csv  A2     A2    B3     J18    C10021.csv   A2     A2    B3     J18    --
C10022.csv  A2     A3    B4     K26    C10022.csv   A2     A3    B4     K26    --
C10023.csv  A2     A2    B3     F16    C10023.csv   A2     A2    B3     F16    --
C10024.csv  A5     A5    B7     K28    C10024.csv   A4     A5    B7     K28    ERROR
C10025.csv  A4     A4    B5     F26    C10025.csv   A4     A4    B5     F26    --
C10026.csv  A4     A5    B6     E24    C10026.csv   A4     A5    B6     E24    --
C10027.csv  A4     A4    B5     F22    C10027.csv   A4     A4    B5     F22    --
  …
```

Category extraction by factorization !

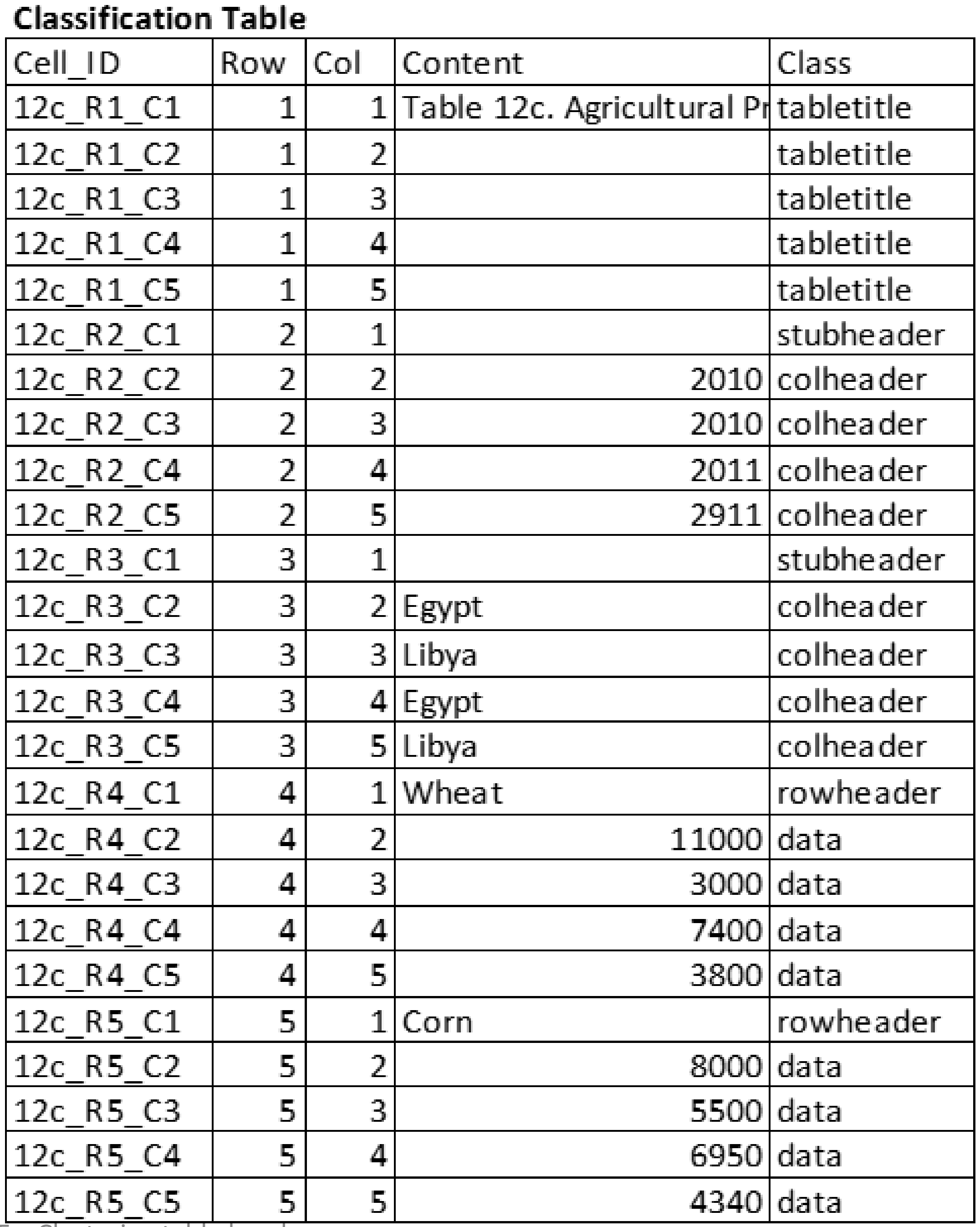
Table 12b, row header (transposed):(Egypt*2010)+(Egypt*2011)+(Libya*2010)+(Libya*2011)
= (Egypt+Libya)*(2010+2011)

Table 12c, column header: (2010*Egypt)+(2010*Libya)+(2011*Egypt)+(2011*Libya)
= (2010+2011)*(Egypt +Libya)

Table 12d row header (transposed): (Wheat*2010)+(Wheat*2011)+(Corn*2010)+(Corn *2011)
= (Wheat + Corn)*(2010+2011)

(*ACM EIA 2011*)

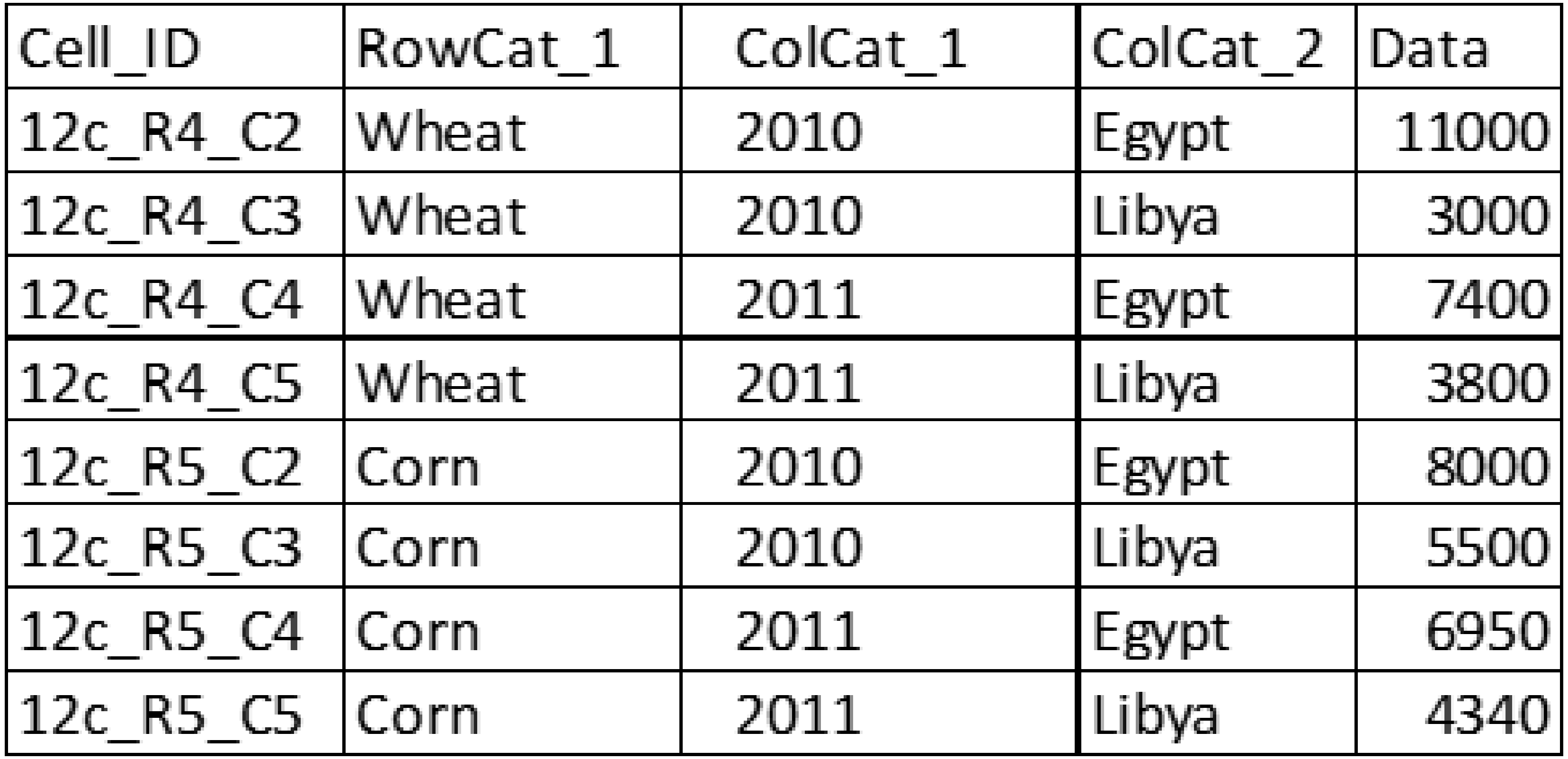
Classification output file



Classification Table

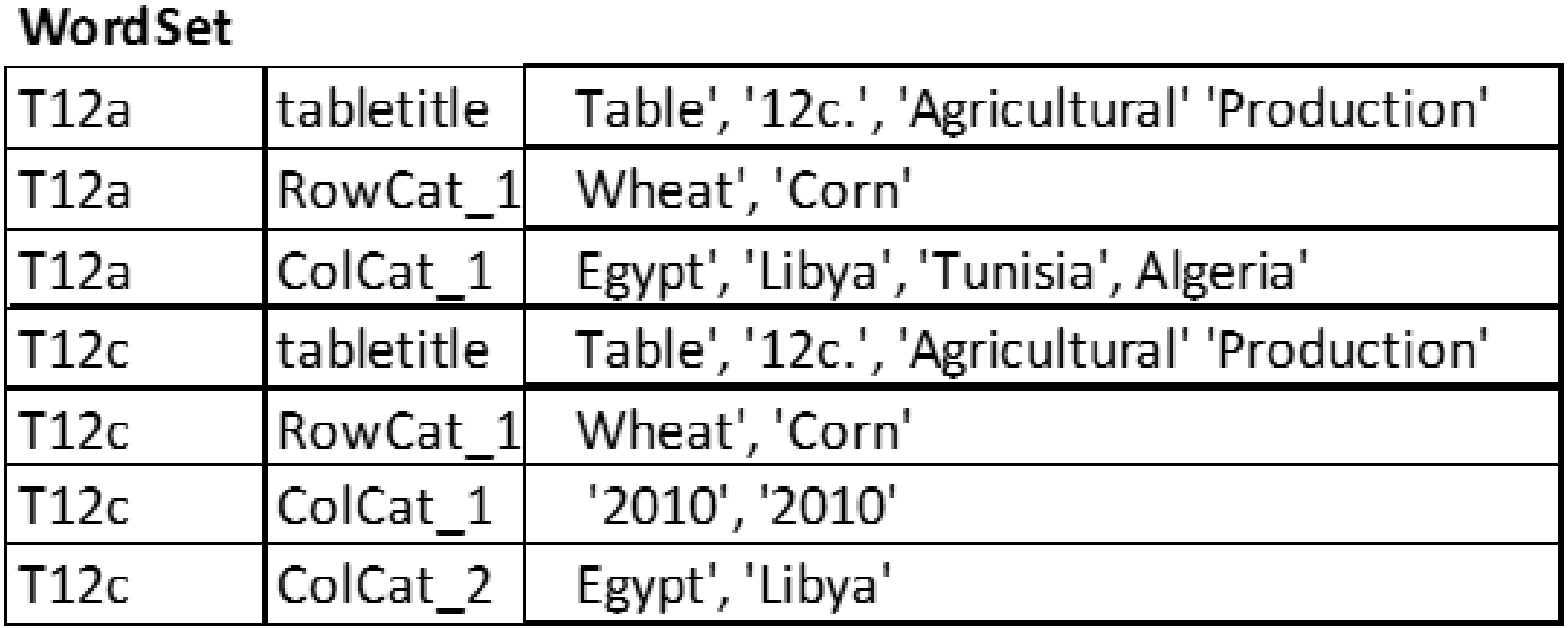
Cell_ID	Row	Col	Content	Class
12c_R1_C1	1	1	Table 12c. Agricultural Pr	tabletitle
12c_R1_C2	1	2		tabletitle
12c_R1_C3	1	3		tabletitle
12c_R1_C4	1	4		tabletitle
12c_R1_C5	1	5		tabletitle
12c_R2_C1	2	1		stubheader
12c_R2_C2	2	2	2010	colheader
12c_R2_C3	2	3	2010	colheader
12c_R2_C4	2	4	2011	colheader
12c_R2_C5	2	5	2911	colheader
12c_R3_C1	3	1		stubheader
12c_R3_C2	3	2	Egypt	colheader
12c_R3_C3	3	3	Libya	colheader
12c_R3_C4	3	4	Egypt	colheader
12c_R3_C5	3	5	Libya	colheader
12c_R4_C1	4	1	Wheat	rowheader
12c_R4_C2	4	2	11000	data
12c_R4_C3	4	3	3000	data
12c_R4_C4	4	4	7400	data
12c_R4_C5	4	5	3800	data
12c_R5_C1	5	1	Corn	rowheader
12c_R5_C2	5	2	8000	data
12c_R5_C3	5	3	5500	data
12c_R5_C4	5	4	6950	data
12c_R5_C5	5	5	4340	data

Canonical Table for Table 12c. This is a relational table that can be read directly into Access or into an a collection of RDF triples for query formulation. !



Cell_ID	RowCat_1	ColCat_1	ColCat_2	Data
12c_R4_C2	Wheat	2010	Egypt	11000
12c_R4_C3	Wheat	2010	Libya	3000
12c_R4_C4	Wheat	2011	Egypt	7400
12c_R4_C5	Wheat	2011	Libya	3800
12c_R5_C2	Corn	2010	Egypt	8000
12c_R5_C3	Corn	2010	Libya	5500
12c_R5_C4	Corn	2011	Egypt	6950
12c_R5_C5	Corn	2011	Libya	4340

WordSet (of unique words) of the table titles and category headers of Table 12a and 12c



WordSet

T12a	tabletitle	Table', '12c.', 'Agricultural' 'Production'
T12a	RowCat_1	Wheat', 'Corn'
T12a	ColCat_1	Egypt', 'Libya', 'Tunisia', Algeria'
T12c	tabletitle	Table', '12c.', 'Agricultural' 'Production'
T12c	RowCat_1	Wheat', 'Corn'
T12c	ColCat_1	'2010', '2010'
T12c	ColCat_2	Egypt', 'Libya'

The Jaccard distance between word sets p and q

$$D_J(p,q) = 1 - |p\cap q| / |p\cup q|$$

D_J is a proper metric:

- $D_J(p,p) = 0$;
- $D_J(p,q) = D_J(q,p)$;
- $0 \leq D_J(p,q)$;
- $D_J(p,r) \leq D_J(p,q) + D_J(q,r)$.

The simplest sequential similarity clustering algorithm

(cf. Hall 1966, *Leader-follower* Adolfio & Samet, CACM Oct 2014 NewsStand)

Input: header and title samples, Θ_{LOW}, Θ_{HIGH}

Samples S_j, j = 1 to m, Clusters C_k, k = 1 to n.

Initialization: S_1 → C_1, n = 1, D_{min} = 1

```
For k = 2 to m                                        # for every sample, in some preset order
    For c = 1 to n                                    # in every cluster
        For j = 1 to |Cc|                             # check every member
            If $D_J(S_k, S_{i(j))} < D_{min}$,
                $D_{min} = D_J(S_k, S_{i(j)})$ and C =c          # keep track of cluster with nearest sample

            If $D_{min}$, < $\Theta_{LOW}$, $S_k$ → $C_C$;              # assign sample to cluster with nearest sample

            If $D_{min}$, > $\Theta_{HIGH}$, $S_k$ → $C_{n+1}$; n → n+1  # or create a new cluster with only this sample
```

Experiment

200 web tables from government sites in six countries

Ground truth only for segmentation (four critical cells)

197 tables correctly segmented

2 tables had duplicate columns

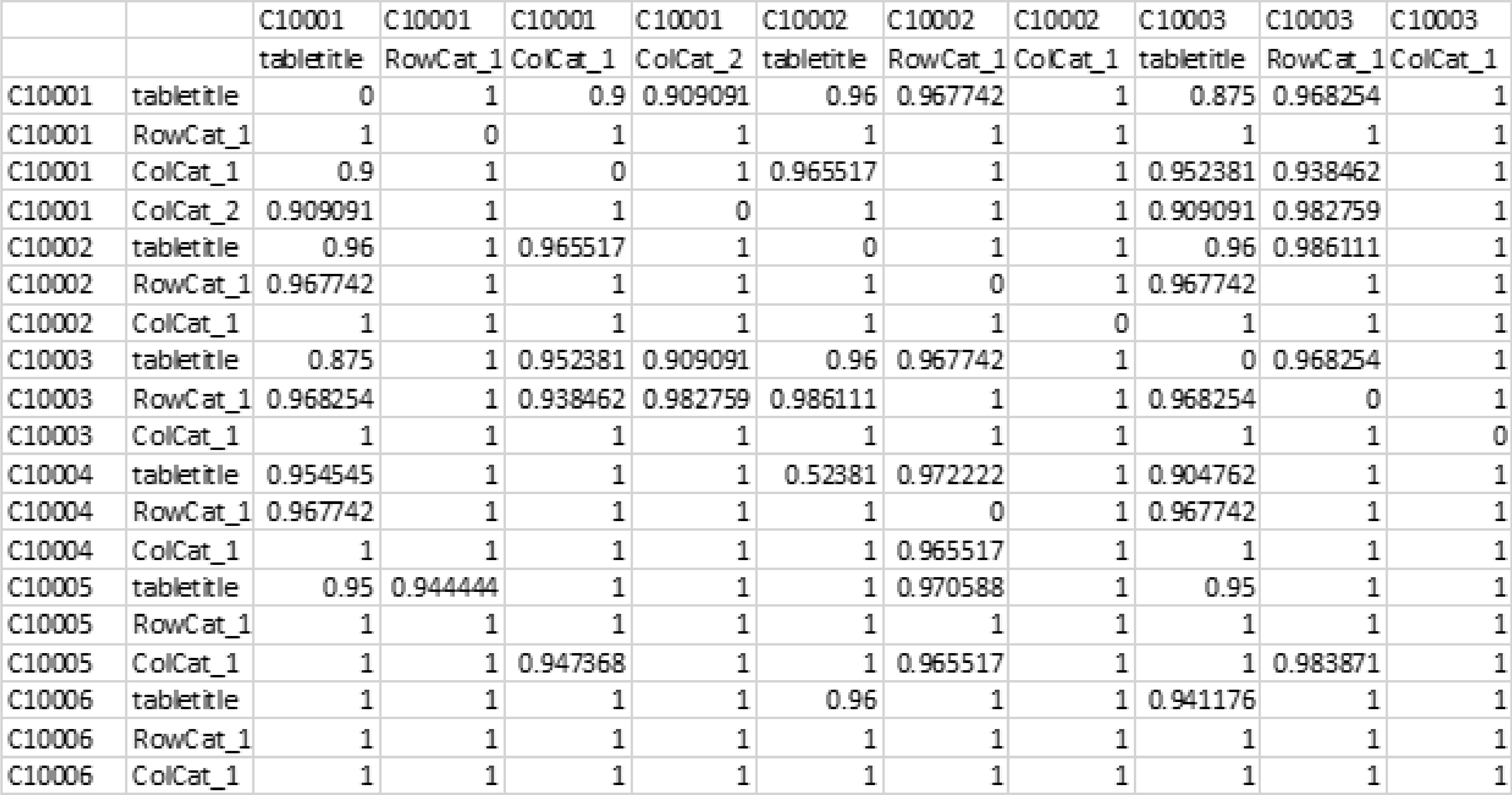
1 table with arguable ground truth

Word sets extracted from the classification and the canonical tables.

615 table titles, row headers and column headers clustered
with various threshold values over the 378,225 computed distances.

217 pairs with distance = 0;   138,393 pairs with distance = 1.

Result: Distance Table (partial output)



		C10001	C10001	C10001	C10001	C10002	C10002	C10002	C10003	C10003	C10003
		tabletitle	RowCat_1	ColCat_1	ColCat_2	tabletitle	RowCat_1	ColCat_1	tabletitle	RowCat_1	ColCat_1
C10001	tabletitle	0	1	0.9	0.909091	0.96	0.967742	1	0.875	0.968254	1
C10001	RowCat_1	1	0	1	1	1	1	1	1	1	1
C10001	ColCat_1	0.9	1	0	1	0.965517	1	1	0.952381	0.938462	1
C10001	ColCat_2	0.909091	1	1	0	1	1	1	0.909091	0.982759	1
C10002	tabletitle	0.96	1	0.965517	1	0	1	1	0.96	0.986111	1
C10002	RowCat_1	0.967742	1	1	1	1	0	1	0.967742	1	1
C10002	ColCat_1	1	1	1	1	1	1	0	1	1	1
C10003	tabletitle	0.875	1	0.952381	0.909091	0.96	0.967742	1	0	0.968254	1
C10003	RowCat_1	0.968254	1	0.938462	0.982759	0.986111	1	1	0.968254	0	1
C10003	ColCat_1	1	1	1	1	1	1	1	1	1	0
C10004	tabletitle	0.954545	1	1	1	0.52381	0.972222	1	0.904762	1	1
C10004	RowCat_1	0.967742	1	1	1	1	0	1	0.967742	1	1
C10004	ColCat_1	1	1	1	1	1	0.965517	1	1	1	1
C10005	tabletitle	0.95	0.944444	1	1	1	0.970588	1	0.95	1	1
C10005	RowCat_1	1	1	1	1	1	1	1	1	1	1
C10005	ColCat_1	1	1	0.947368	1	1	0.965517	1	1	0.983871	1
C10006	tabletitle	1	1	1	1	0.96	1	1	0.941176	1	1
C10006	RowCat_1	1	1	1	1	1	1	1	1	1	1
C10006	ColCat_1	1	1	1	1	1	1	1	1	1	1

Result: Cluster membership vs. thresholds

θ_{LOW}	**0.00**	**0.05**	**0.05**	**0.50**
θ_{HIGH}	**1.00**	**0.95**	**0.05**	**0.50**
Number of multi-member clusters	9	11	52	72
Samples in multi-member clusters	33	49	155	290
Number of single-member clusters	50	86	460	325

Values not sensitive to random permutations of order of presentation

Result: Example of program output

```
ClusterTable__7_27_2014_16h28m.csv

C10001_RowCat_1    C10008_ColCat_1     C10073_RowCat_1    C10080_ColCat_1

2008
2007
2006
2005
2004
2003
2002
```

The program found two tables with identical row *and* column headers. A duplicate! Another possible use.

Observations

Stop words:

4.1% of category headers and 20.1% of table titles

Synonyms:

Found for 46.6% of category headers and 55.1% of table titles
Examples: 2: two, deuce, US state names: hi, me, in, or, ok

Queries:

Executed in Access, Virtuoso and Protégé,
but none yet making use of clustering results

Incremental contribution:

A scalable measure of table similarity
(12s for segmentation and classification, + 3s for clustering)

Observations

Stop words:

4.1% of category headers and 20.1% of table titles

Synonyms:

Found for 46.6% of category headers and 55.1% of table titles
Examples: 2: two, deuce, US state names: hi, me, in, or, ok

Queries:

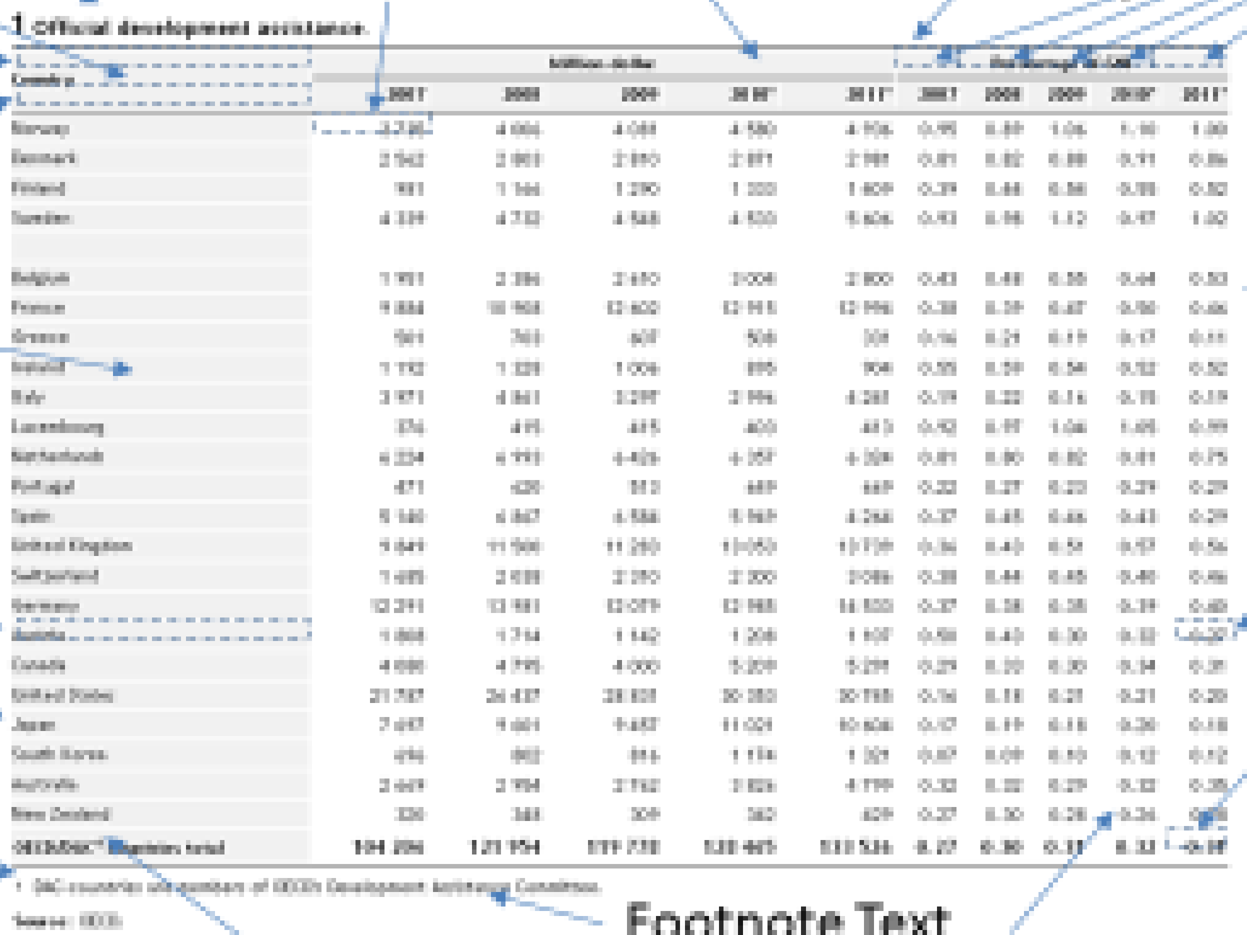
Executed in Access, Virtuoso and Protégé,
but none yet making use of clustering results

Incremental contribution:

A scalable measure of table similarity
(12s for segmentation and classification, + 3s for clustering)

THANK YOU!

Table Terminology and Critical Cells (CCs)


1 Official development assistance

Stub Header

Table title

CC3

Column Header

a Spanning Cell

Spanning Cell Components

CC1

CC2

Row Header or Stub

Data Region

an Atomic Header Cell

a Data Cell

a Header Label

CC4

Footnote Prefix

Footnote Text

Footnote Marker

a Data Value