Fun with data linkage: SQL join versus SAS merge

Christine Werk

Contributors: the CYDL team and partners

Child and Youth Data Laboratory (CYDL)

ALBERTA CENTRE FOR
CHILD, FAMILY
& COMMUNITY
RESEARCH

The Alberta Centre for Child, Family and Community Research (ACCFCR)seeks to improve child well being by conducting, funding and mobilizing research for evidence-informed policy and practice.

The ACCFCR manages the Child and Youth Data Laboratory (CYDL).

Researchers at the CYDL link and analyze administrative data from Government of Alberta ministries.

CYDL's Cross-Sectional Project

714,000 youth 12 to 24 years of age

5 ministries providing data

11 databases

2008/09 service use

CYDL's Longitudinal Project

About 2 million people 0 to 30 years of age

5 ministries providing data

23 databases; more than 225 data elements

2005/06 to 2010/11 service use

Ministries

Health
Education
Innovation and Advanced Education
Human Services
Justice and Solicitor General

Processing Justice data

Clean data

Match to offence library (Statistic Canada)

Compute most serious offence

Join most serious offence categories with other ministry data

Person A

Raw JOIN data
CC 430(1.1)(A)
CDSA 546.33
YCJA 899

+

Offence library
01 430 1.1 A
22 546.22
24 899

MSO category
Property offence

Person A

JOIN data
Property offence
Administrative offences
Number of charges
Reoffending

+

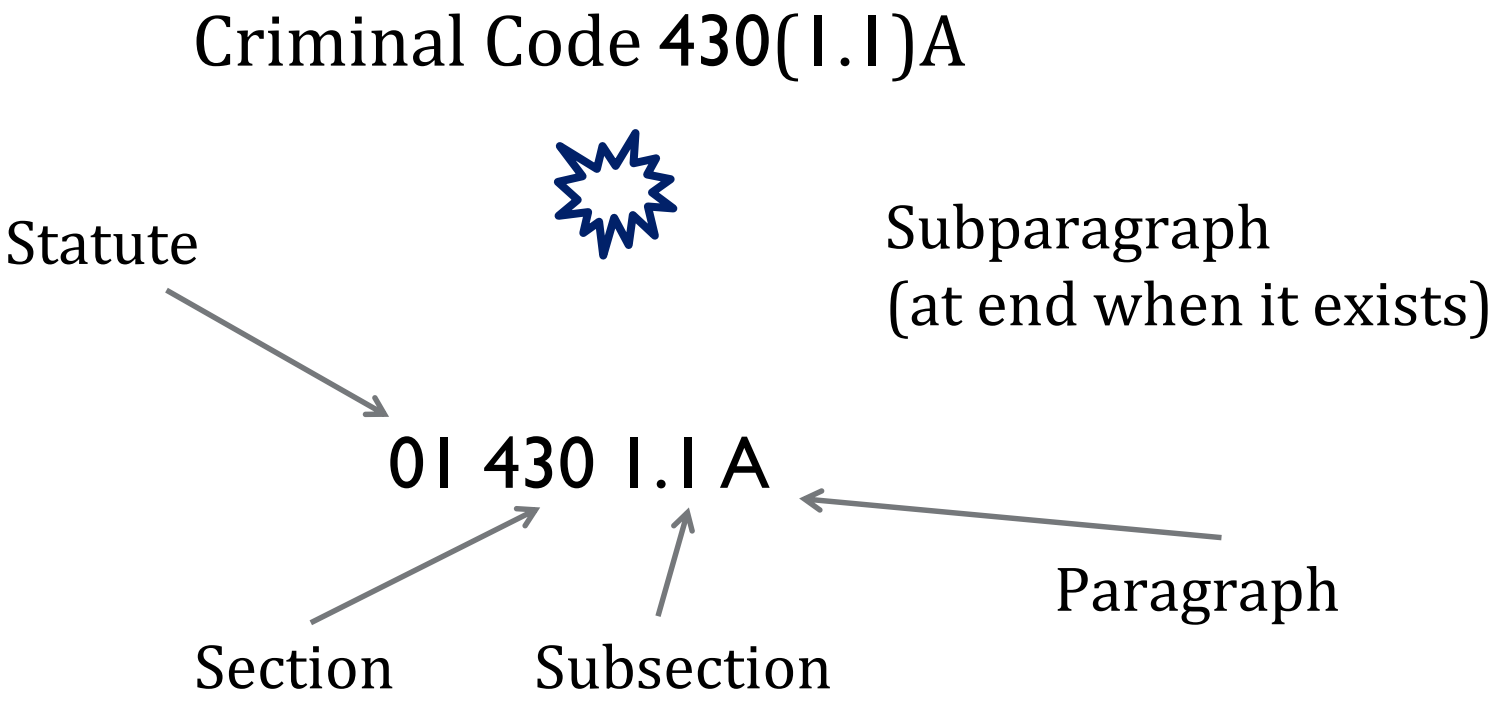
Other data
Meeting expectations
Visited an emergency room
Aboriginal status

CYDL Report

Statistics Canada Offence Library

Charge code: Statute, section, subsection, paragraph, subparagraph
Date: Active date, inactive date
Category grouper
Severity index

Example justice code



Most Serious Offence

Match on charge code and date

Determine a category for each charge

Choose the most serious for each person

Challenge….

Match join codes to offence library by all components of the charge code as well as date

Statute + Section + Subsection + Paragraph + Subparagraph +
Offence date (between active and inactive dates)

Justice data		Offence library
Statue	→	Statue
Section	→	Section
Subsection	→	Subsection
Paragraph	→	Paragraph
Subparagraph	→	Subparagraph
Offence date	→	Date active
		Date inactive

SAS merge

MERGE SAS-data-set-list;
BY variable-list;
SAS-data-set-list
variable-list

Sorting

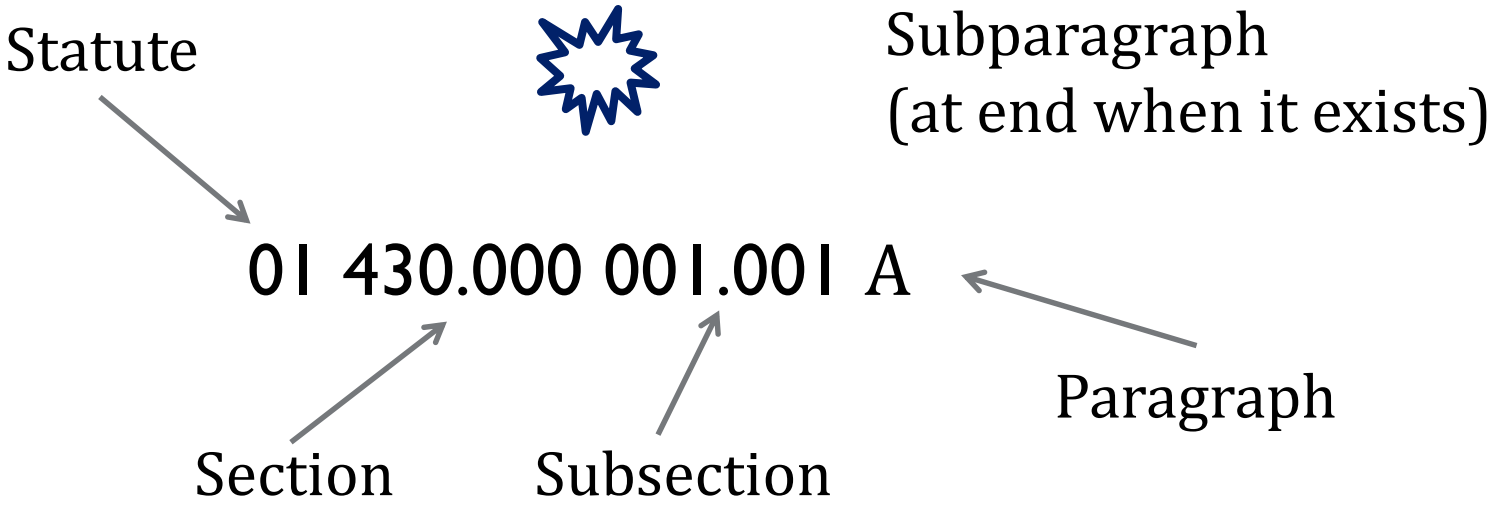
Criminal Code 430(1.1)A, 380B, 226(3)

01	430	1.1	A
01	380		B
01	226	3	

01	226		
01	380	1.1	A
01	430	3	B

Zero padding

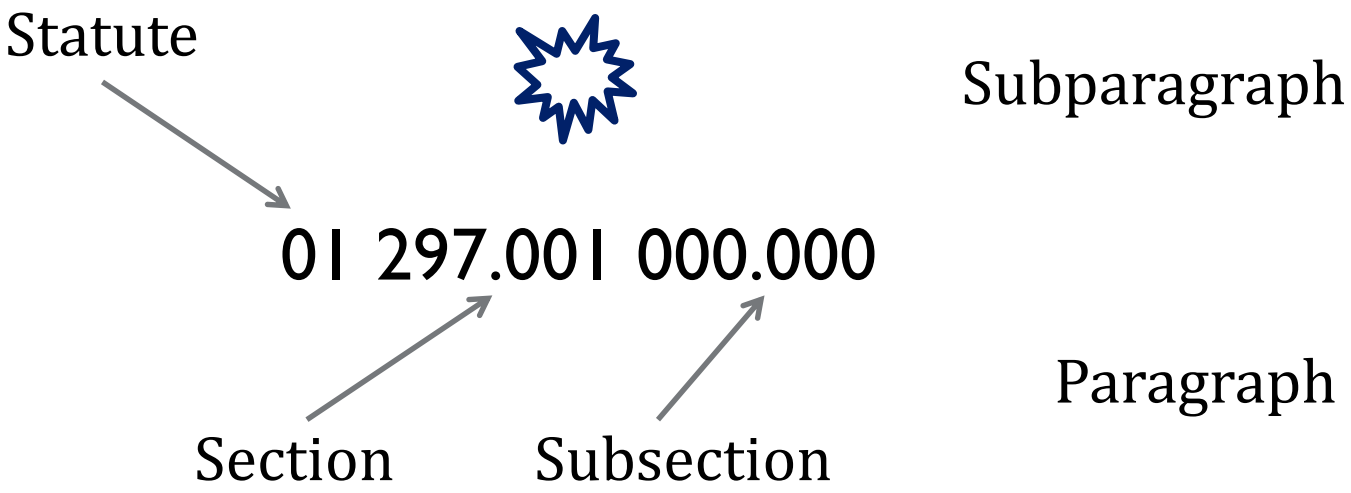
Criminal Code 430(1.1)A



Zero padding…

Criminal Code 297.01

Criminal Code 297.1



SAS merge

Not possible unless zero padding used
Zero padding has challenges

Other option??

SQL join

```
SELECT column_name(s)
FROM table1
JOIN table2
ON table1.column_name=table2.column_name;
```

SQL join

Benefits: do not need to sort, can merge on multiple variables

Challenges: learning new code structure, takes a bit longer to run

Solution!!

```
PROC SQL;
CREATE TABLE OffenceMatch AS
SELECT P. *, O.date_active, O.date_inactive, O.COC,
O.index_s
FROM P2MSO P LEFT JOIN  OffenceLib O
ON P.statute = O.statute AND
P.section = O.section AND
P.subsection = O.subsection AND
P.paragraph = O.paragraph AND
P.subparagraph = O.subparagraph
WHERE (P.offence_date GE O.date_active AND
P.offence_date LE O.date_inactive)
ORDER BY P.CYDLID; QUIT;
```

Cool features of SQL

P. and O. short hand
No need to sort by each column
Single step of code
Reduced use of semicolon key

Contributors

Partnering ministries

Aboriginal Relations
Education
Enterprise and Advanced Education
Health
Human Services
Justice and Solicitor General

Committees and Working Groups

The Deputy Ministers Committee
The Assistant Deputy Ministers Committee
The Coordinating Committee
The Research Working Group
The Data/Technical Working Group
The Legal/Privacy Working Group

The Alberta Centre for Child, Family and Community Research

CYDL Research Team
Xinjie Cui, PhD
Leslie Twilley, PhD
Cecilia Bukutu, PhD
Christine Werk, PhD
Navjot Lamba, PhD
Ozlem Cankaya, PhD
Saiful Kabir, MHI
Hitesh Bhatt, MSc
Hesam Izakian
ACCFCR Scientific Director
Suzanne Tough, PhD

And many other reviewers and contributors

For more information, go to research4children.com and click on CYDL

A	B	C	D	E	F	G	H	I	J	K	L	M
STATUTE	SECTION	SUBSECT	PARA	SUBPARA	DTACTIVE	DTINACTIVE	PROCEDUR	UCR2	COC	MSOCategory	RANK	
1	429	3			December 12, 1988	January 1, 2099	0	3810	26	Other CC	176	
1	429	.			December 12, 1988	January 1, 2099	0	3810	26	Other CC	176	
1	430	1	A		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1	B		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1	C		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1	D		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1			December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1.1	A		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1.1	B		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1.1	C		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1.1	D		December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	1.1			December 12, 1988	January 1, 2099	0	2170	15	Property	156	
1	430	2			December 12, 1988	January 1, 2099	0	1630	15	Property	75	
1	430	3	A		December 12, 1988	February 14, 1995	0	2172	15	Property	157	
1	430	3	A		February 15, 1995	January 1, 2099	0	2172	15	Property	157	
1	430	3	B		December 12, 1988	February 14, 1995	0	2172	15	Property	157	
1	430	3	B		February 15, 1995	January 1, 2099	0	2172	15	Property	157	