Working with pandas DataFrames

Presenter: Steve Baskauf
vanderbi.lt/codegraf

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

CodeGraf landing page

- vanderbi.lt/codegraf

pandas DataFrame

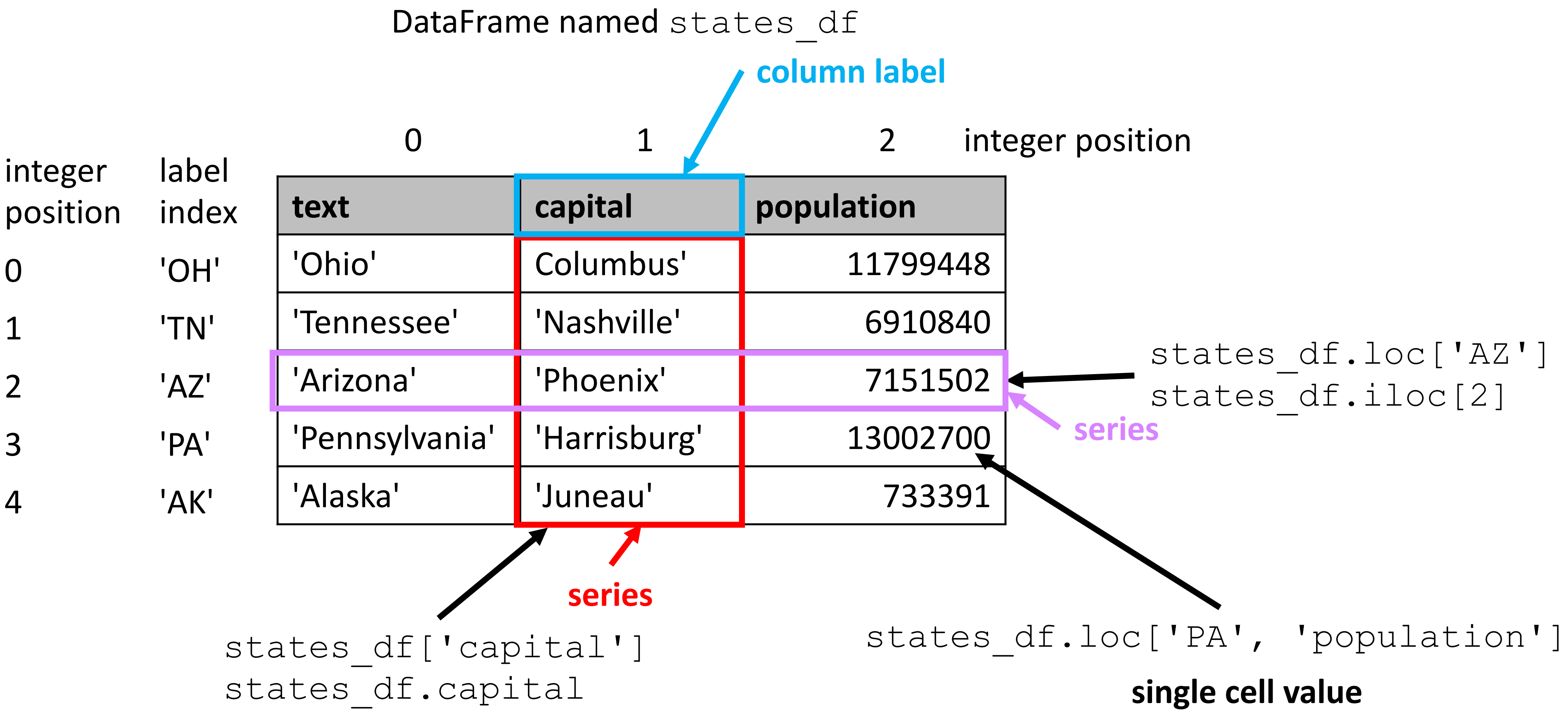
DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

DataFrame

- A specific two-dimensional data structure designed to be like a **spreadsheet**.
- pandas DataFrames can be built from a **set of Series** (one Series for each column).
- Column series share a **common label index**.
- Data frames are built by instantiating a **`pd.DataFrame()`** object.

DataFrame



DataFrame named states_df

column label

		0	1	2
integer position	label index	text	capital	population
0	'OH'	'Ohio'	Columbus'	11799448
1	'TN'	'Tennessee'	'Nashville'	6910840
2	'AZ'	'Arizona'	'Phoenix'	7151502
3	'PA'	'Pennsylvania'	'Harrisburg'	13002700
4	'AK'	'Alaska'	'Juneau'	733391

integer position

```
states_df.loc['AZ']
states_df.iloc[2]
```

series

series

```
states_df['capital']
states_df.capital
```

```
states_df.loc['PA', 'population']
```

single cell value

- **Data frames** are essentially tables.
- The **values** of columns or rows are **series.**

Data frame indexing and attributes

`.loc[]` to refer to a row by label index

`.iloc[]` to refer to a row by integer position

`.columns` to refer to the column labels

`.index` to refer to the line label indices

Loading a DataFrame from a file

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

File read and write functions

`pd.read_csv()` read from a CSV file into a data frame.

`pd.to_csv()` write from a data frame to a CSV file.

`pd.read_excel()` read from an Excel file into a data frame.

`pd.to_excel()` write from a data frame to an Excel file.

The read functions can be performed using a web URL

Data frame attributes and methods

`.head()` to display first five lines

`.tail()` to display last five lines

`.shape` returns a tuple of (number_rows, number_columns)

Data types and missing data

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Data types for DataFrames from files

- By default, Pandas guesses for a column based on the values in it.
- Use the **`dtype=str`** argument to force all cells to be strings.

Missing data

- By default, empty cells are given the NumPy missing data value: **`NaN`**
- Use the **`na_filter=False`** argument to read in empty cells as empty strings.
- **`NaN`** is considered a number, so turning off the **`na_filter`** can result in mixed data columns (type: object)

Setting the label index

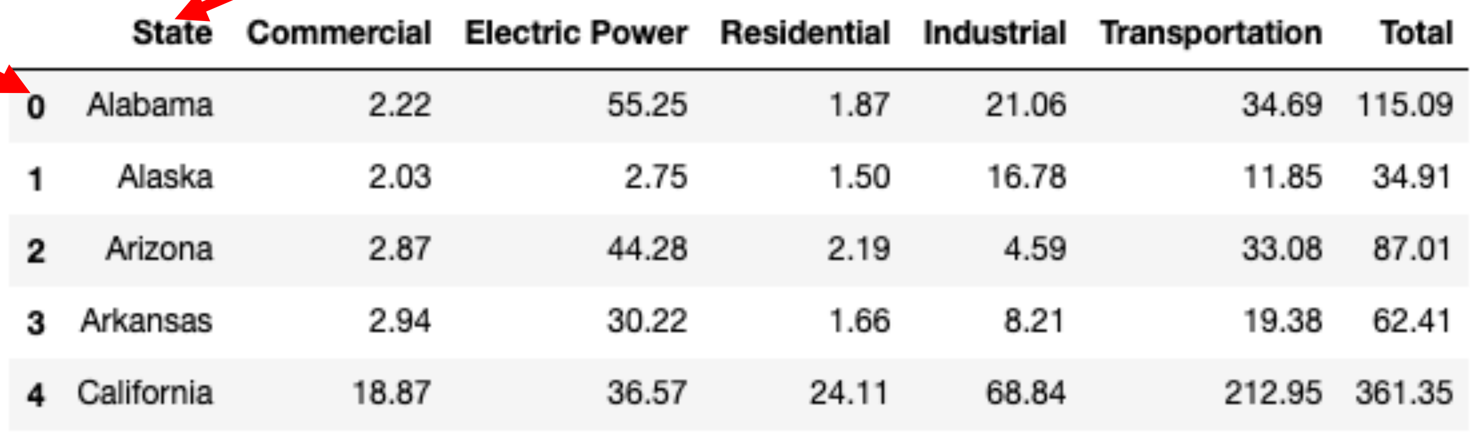
DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Column vs. index label

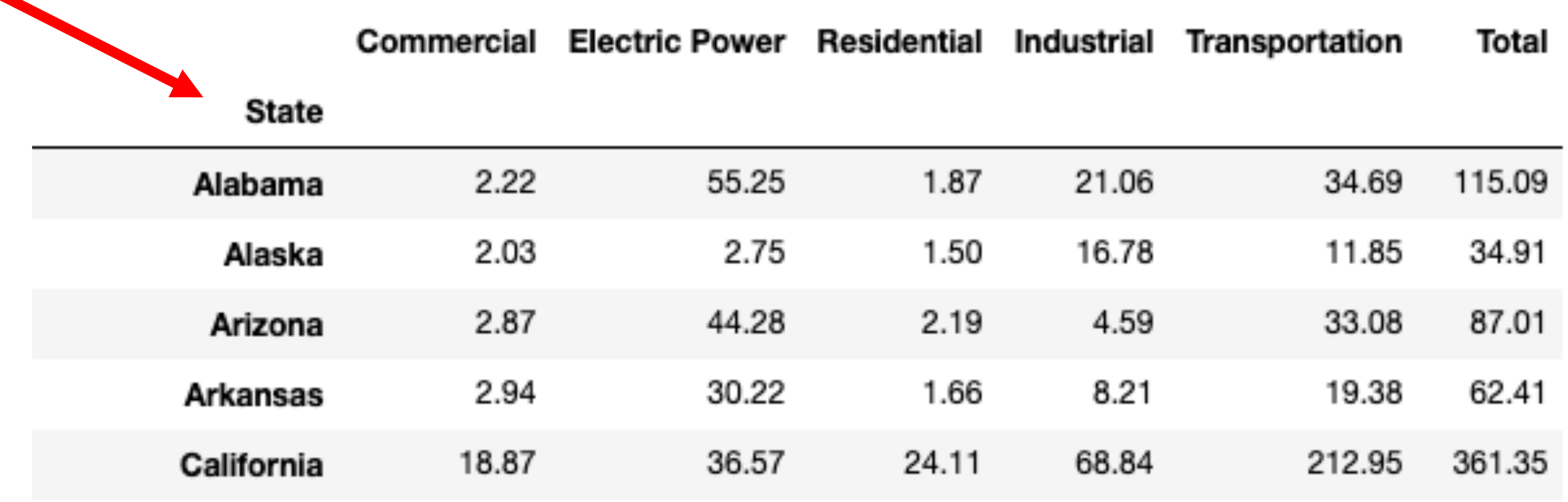
integer label index

regular column



	State	Commercial	Electric Power	Residential	Industrial	Transportation	Total
0	Alabama	2.22	55.25	1.87	21.06	34.69	115.09
1	Alaska	2.03	2.75	1.50	16.78	11.85	34.91
2	Arizona	2.87	44.28	2.19	4.59	33.08	87.01
3	Arkansas	2.94	30.22	1.66	8.21	19.38	62.41
4	California	18.87	36.57	24.11	68.84	212.95	361.35

string label index



	Commercial	Electric Power	Residential	Industrial	Transportation	Total
State						
Alabama	2.22	55.25	1.87	21.06	34.69	115.09
Alaska	2.03	2.75	1.50	16.78	11.85	34.91
Arizona	2.87	44.28	2.19	4.59	33.08	87.01
Arkansas	2.94	30.22	1.66	8.21	19.38	62.41
California	18.87	36.57	24.11	68.84	212.95	361.35

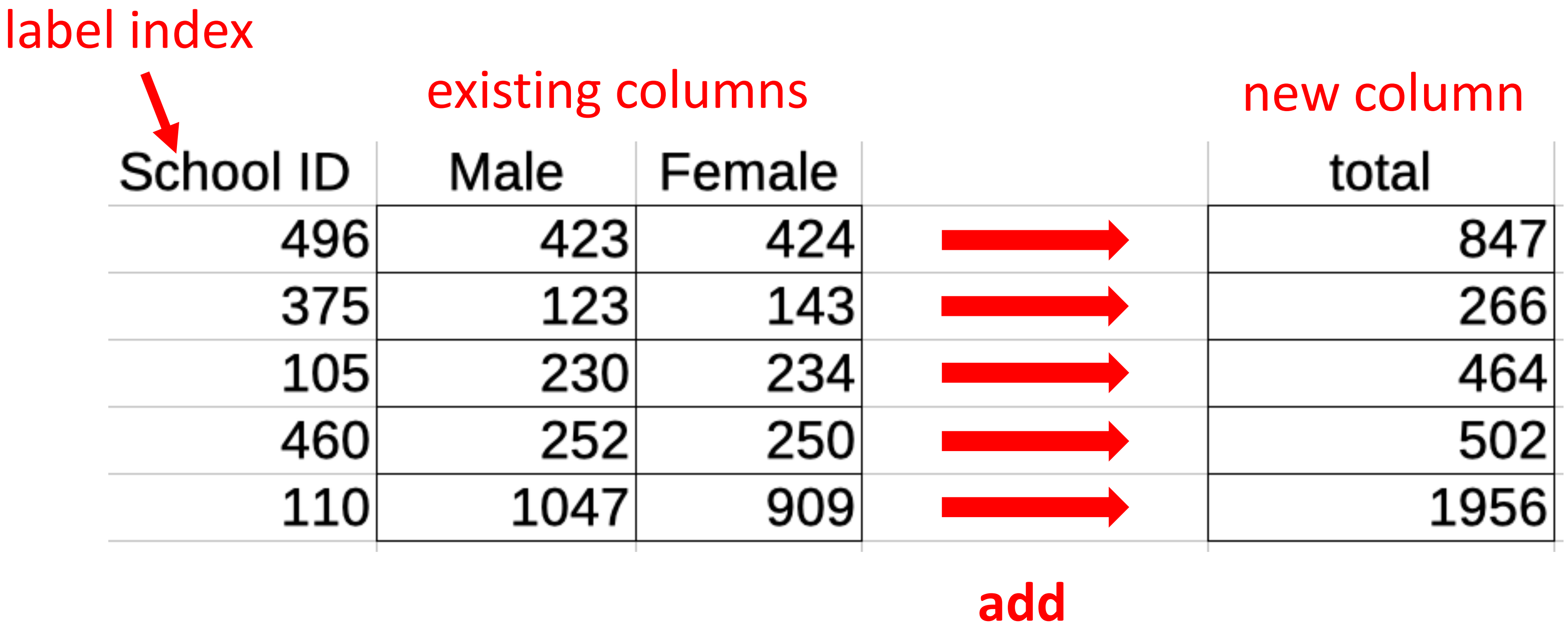
Creating a calculated column

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Adding two columns as a vectorized operation

```
schools_df['total'] = schools_df['Male'] + schools_df['Female']
```



label index

existing columns

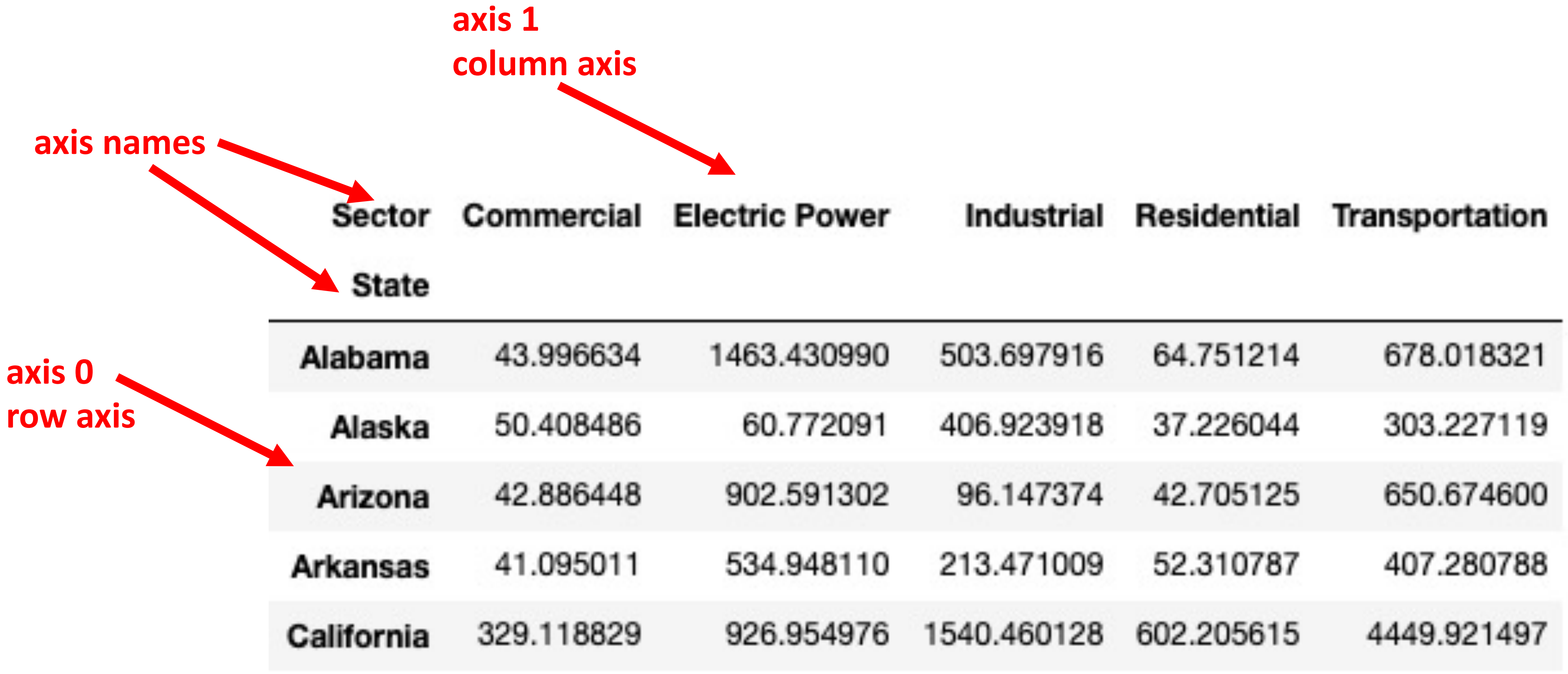
new column

School ID	Male	Female	total
496	423	424	847
375	123	143	266
105	230	234	464
460	252	250	502
110	1047	909	1956

add

Simple column methods

"Axes" of a data frame



axis 1
column axis

axis names

axis 0
row axis

Sector	Commercial	Electric Power	Industrial	Residential	Transportation
State					
Alabama	43.996634	1463.430990	503.697916	64.751214	678.018321
Alaska	50.408486	60.772091	406.923918	37.226044	303.227119
Arizona	42.886448	902.591302	96.147374	42.705125	650.674600
Arkansas	41.095011	534.948110	213.471009	52.310787	407.280788
California	329.118829	926.954976	1540.460128	602.205615	4449.921497

For more videos like this, visit the CodeGraf landing page

vanderbi.lt/codegraf

DiSC DIGITAL SCHOLARSHIP
AND COMMUNICATIONS

VANDERBILT
UNIVERSITY

Dropping and transposing

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

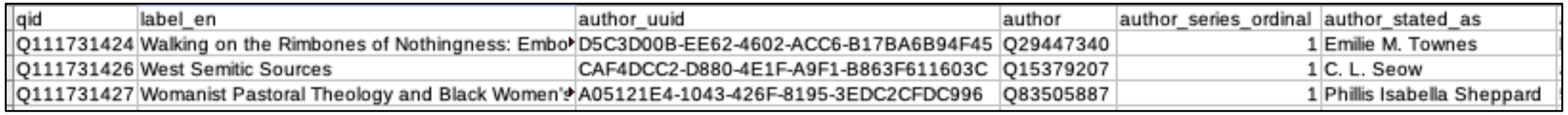
Jean & Alexander Heard LIBRARIES

Combining DataFrames

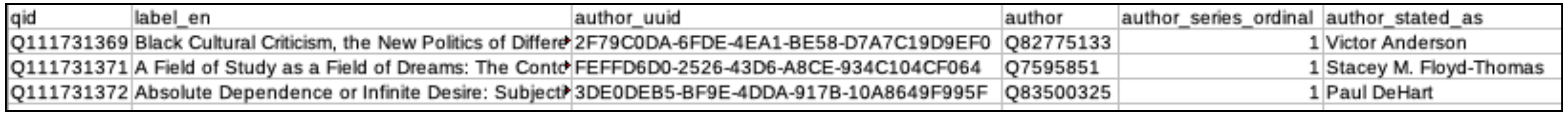
DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

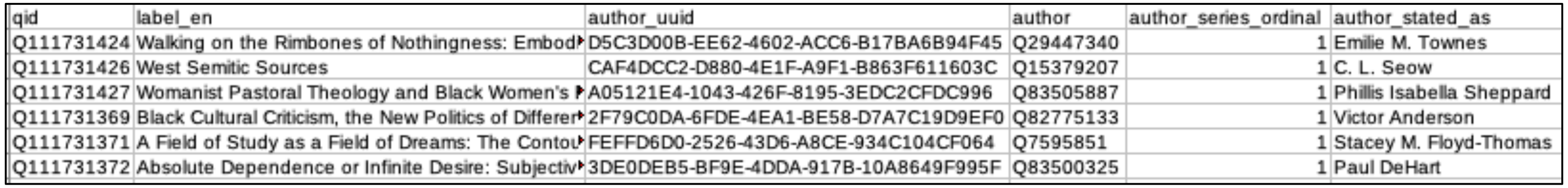
Concatenating DataFrames



qid	label_en	author_uuid	author	author_series_ordinal	author_stated_as
Q111731424	Walking on the Rimbones of Nothingness: Embo	D5C3D00B-EE62-4602-ACC6-B17BA6B94F45	Q29447340	1	Emilie M. Townes
Q111731426	West Semitic Sources	CAF4DCC2-D880-4E1F-A9F1-B863F611603C	Q15379207	1	C. L. Seow
Q111731427	Womanist Pastoral Theology and Black Women's	A05121E4-1043-426F-8195-3EDC2CFDC996	Q83505887	1	Phillis Isabella Sheppard



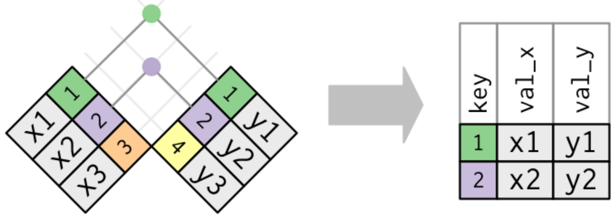
qid	label_en	author_uuid	author	author_series_ordinal	author_stated_as
Q111731369	Black Cultural Criticism, the New Politics of Differe	2F79C0DA-6FDE-4EA1-BE58-D7A7C19D9EF0	Q82775133	1	Victor Anderson
Q111731371	A Field of Study as a Field of Dreams: The Conto	FEFFD6D0-2526-43D6-A8CE-934C104CF064	Q7595851	1	Stacey M. Floyd-Thomas
Q111731372	Absolute Dependence or Infinite Desire: Subject	3DE0DEB5-BF9E-4DDA-917B-10A8649F995F	Q83500325	1	Paul DeHart



qid	label_en	author_uuid	author	author_series_ordinal	author_stated_as
Q111731424	Walking on the Rimbones of Nothingness: Embod	D5C3D00B-EE62-4602-ACC6-B17BA6B94F45	Q29447340	1	Emilie M. Townes
Q111731426	West Semitic Sources	CAF4DCC2-D880-4E1F-A9F1-B863F611603C	Q15379207	1	C. L. Seow
Q111731427	Womanist Pastoral Theology and Black Women's	A05121E4-1043-426F-8195-3EDC2CFDC996	Q83505887	1	Phillis Isabella Sheppard
Q111731369	Black Cultural Criticism, the New Politics of Differer	2F79C0DA-6FDE-4EA1-BE58-D7A7C19D9EF0	Q82775133	1	Victor Anderson
Q111731371	A Field of Study as a Field of Dreams: The Contou	FEFFD6D0-2526-43D6-A8CE-934C104CF064	Q7595851	1	Stacey M. Floyd-Thomas
Q111731372	Absolute Dependence or Infinite Desire: Subjectiv	3DE0DEB5-BF9E-4DDA-917B-10A8649F995F	Q83500325	1	Paul DeHart

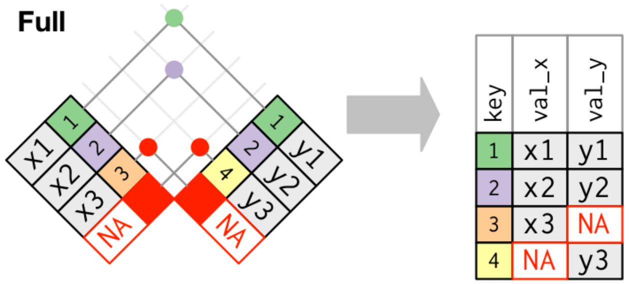
Table joins

- A **join** merges data from two DataFrames
- **`on=`** arguments are the columns used to match table rows
- **Inner join** only outputs rows with matching keys



key	val_x	val_y
1	x1	y1
2	x2	y2

- **Outer join** includes rows that don't match (with **`NaN`** values inserted)



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

- Diagrams from https://r4ds.had.co.nz/relational-data.html

A simple data wrangling workflow

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES