Northeastern University
College *of* Computer and Information Science

CS1100: Computer Science and Its Applications

Text Processing

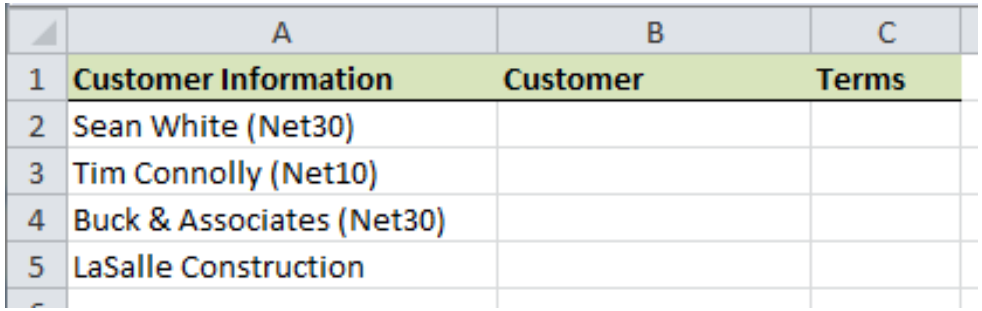
Created By
Martin Schedlbauer
m.schedlbauer@neu.edu

Processing Text

- Excel can be used not only to process numbers, but also text.
- This often involves taking apart (parsing) or putting together text values (strings).
- The parts into which we split a string will be called fields.
- Fields may be separated by delimiting text
- And/or fields may have a fixed width which permits them to be identified.

Example

- Text processing is often necessary when files are imported from other programs:



	A	B	C
1	**Customer Information**	**Customer**	**Terms**
2	Sean White (Net30)		
3	Tim Connolly (Net10)		
4	Buck & Associates (Net30)		
5	LaSalle Construction		

- We'd like to extract the customer name and the payment terms from the text in column A.

Terminology

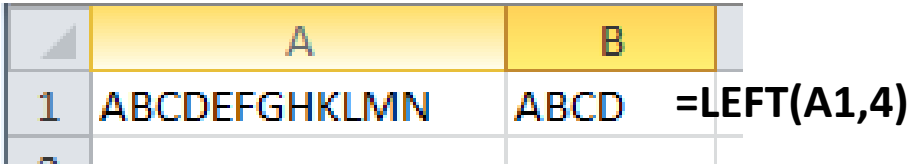
- The process of taking text values apart is called *parsing*.
  - text value = string
  - part of a text value = substring

Text Processing Functions

- Excel provides a number of functions for parsing text:
  - **RIGHT** – take part of the right side of a text value
  - **LEFT** – take part of the left side of a text value
  - **MID** – take a substring within a text value
  - **LEN** – determine the number of characters in a text value
  - **FIND** – find the start of a specific substring within a text value

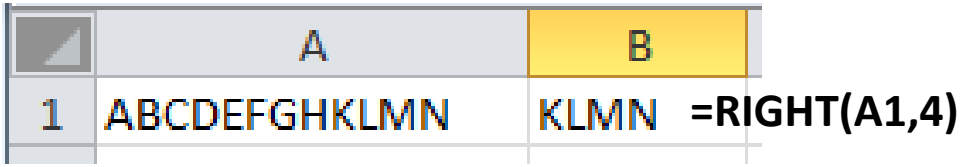
LEFT Function

- The **LEFT** function extracts a specific number of characters from the left side of a text value:



RIGHT Function

- The **RIGHT** function extracts a specific number of characters from the right side (end) of a text value:

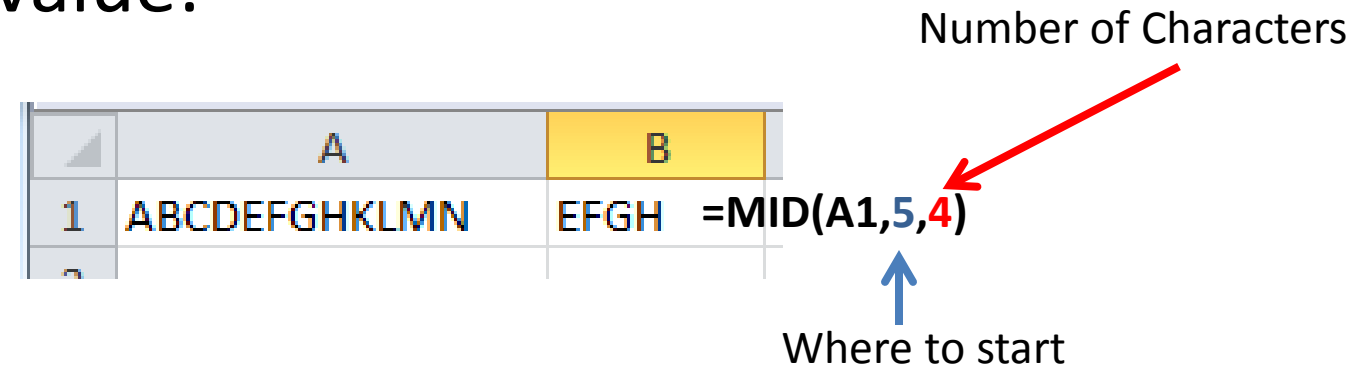


	A	B	
1	ABCDEFGHKLMN	KLMN	=RIGHT(A1,4)

- **SPECIFY THE NUMBER OF CHARACTERS, NOT WHERE TO START!**

MID Function

- The **MID** function extracts some number of characters starting at some position within a text value:



FIND Function

- **FIND** returns the position where a substring starts within a string.
- Finds the first occurrence only.
- Returns a *#VALUE!* error if the substring cannot be found.

	A	B	
1	ABCDEFGHKLMN	4	=FIND("DEF",A1)
2	ABCDEF GHKLMN	7	=FIND(" ",A2)
3	ABCDEF, GHKLMN	7	=FIND(",",A3)

Case Sensitivity

- Note that **FIND** is case sensitive.
- As an alternative, Excel has a **SEARCH** function which is not case sensitive but otherwise works the same way as **FIND**.

16	ABCDEFGHKLMN	#VALUE!	**=FIND("cde",A16)**
17	ABCDEFGHKLMN	3	**=SEARCH("cde",A17)**

IFERROR and **FIND**

- Since **FIND** returns an error when a substring cannot be found, we need to use a sentinel value.

5	ABCDEF, GHKLMN	#VALUE!	**=FIND("[",A5)**

7	ABCDEF, GHKLMN		**=IFERROR(FIND("[",A5),"")**

LEN Function

- The **LEN** function returns the total number of characters in a text, *i.e.*, the "length" of the text value:

9	ABCDEF, GHKLMN	14	**=LEN(A9)**

LEN Function

- The **LEN** function returns the total number of characters in a text, *i.e.*, the "length" of the text value:

- A is the first character
- N is the 14th character

TRIM Function

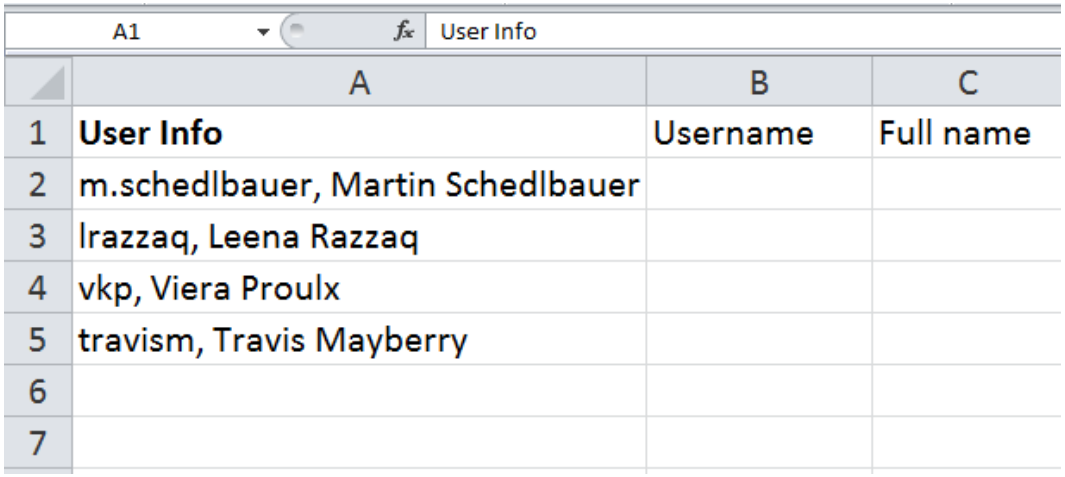
- The **TRIM** function removes all spaces before and after a piece of text. Spaces between words are not removed.
- This is useful if the text you are trying to parse has trailing spaces which may result in errors later
  - For example, if you need to use a result later in a VLOOKUP function.

Example 1 – Delimiting Text

- You are given a list of usernames, each followed by a comma, then a space, then the user's full name
- A comma followed by a space **only** appears between the username and full name
- **Everything** following the username, the comma and the space is the user's full name

Locating the Delimiter (where to split the text)

- The first step is to identify the location where the split will be made
- The split location may be identified by
  - Delimiting text
  - A fixed width field



A1 | User Info

	A	B	C
1	**User Info**	Username	Full name
2	m.schedlbauer, Martin Schedlbauer		
3	lrazzaq, Leena Razzaq		
4	vkp, Viera Proulx		
5	travism, Travis Mayberry		
6			
7			

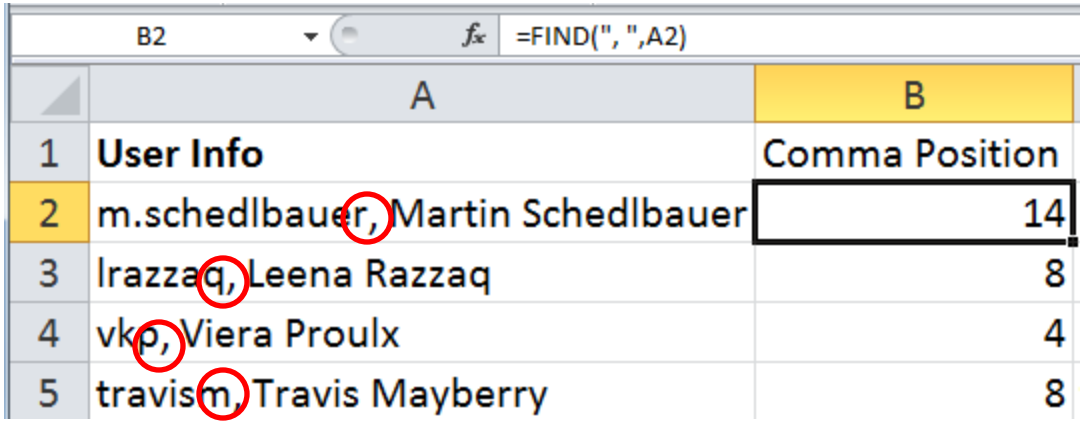
Delimiting Text

- Delimiting text is any sequence of characters that can reliably be used to end one part of the text to be split and the beginning of another.
- In this example, a comma followed by a space can serve as delimiting text.
- On the other hand, the width of each field may vary, so we cannot identify the splitting location by field widths

Finding the Delimiting Text

- Since the width of each field may vary, and we cannot identify the splitting location by field widths, we need to find the location of the comma and space
- Use FIND to return the location of the delimiter.

=FIND(“, ”,A2)



Splitting the Text

- **LEFT**: Number of characters to read
  - Start position = 1
  - End Position = Find(delimiter, cell) – 1
  - Number of characters =

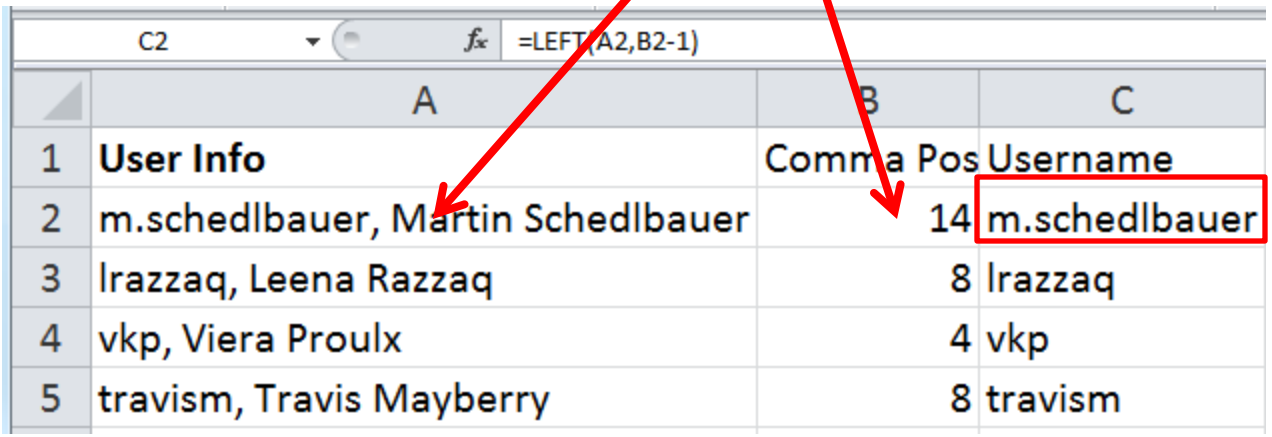
    End position – Start Position + 1 =

    End position

Splitting the Text

- Once we have found the delimiting text, we can split the original text using functions like LEFT, RIGHT and MID
- Note that we must adjust the length in our function to omit the delimiting text.

=LEFT(A2, B2 – 1)



	A	B	C
1	**User Info**	Comma Pos	Username
2	m.schedlbauer, Martin Schedlbauer	14	m.schedlbauer
3	lrazzaq, Leena Razzaq	8	lrazzaq
4	vkp, Viera Proulx	4	vkp
5	travism, Travis Mayberry	8	travism

Splitting the Text

- **RIGHT**: Number of characters to read
  - Start position =

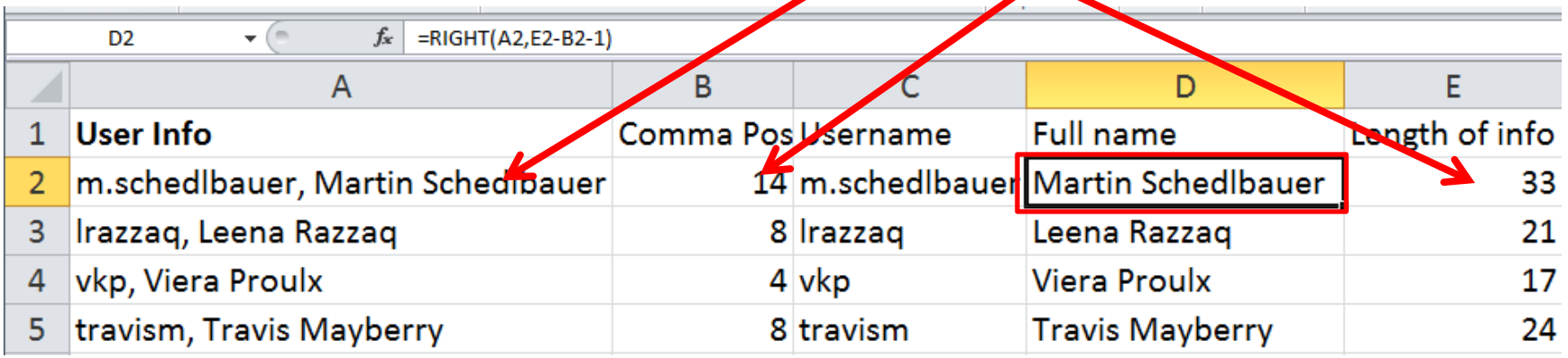
    FIND(delimiter, cell) + LEN(delimiter)
  - End Position = LEN(cell)
  - Number of characters =

    End position – Start Position + 1 =

Splitting the Text

- Using the RIGHT function to find the full name, we need to find the number of characters from the right
  - Subtract the length of the whole text by the location of the delimiter and adjust to omit the delimiter

=RIGHT(A2, E2 – (B2+2) + 1)



	A	B	C	D	E
1	**User Info**	Comma Pos	Username	Full name	Length of info
2	m.schedlbauer, Martin Schedlbauer	14	m.schedlbauer	Martin Schedlbauer	33
3	lrazzaq, Leena Razzaq	8	lrazzaq	Leena Razzaq	21
4	vkp, Viera Proulx	4	vkp	Viera Proulx	17
5	travism, Travis Mayberry	8	travism	Travis Mayberry	24

Splitting the Text

- Using the RIGHT function to find the full name, we need to find the number of characters from the right
  - Subtract the length of the whole text by the location of the delimiter and adjust to omit the delimiter

=RIGHT(A2, E2 – B2 – 1)

D2 =RIGHT(A2,E2-B2-1)

	A	B	C	D	E
1	**User Info**	Comma Pos	Username	Full name	Length of info
2	m.schedlbauer, Martin Schedlbauer	14	m.schedlbauer	Martin Schedlbauer	33
3	lrazzaq, Leena Razzaq	8	lrazzaq	Leena Razzaq	21
4	vkp, Viera Proulx	4	vkp	Viera Proulx	17
5	travism, Travis Mayberry	8	travism	Travis Mayberry	24

Splitting the Text

- **MID**: Start Position, Number of characters to read
  - Start position =

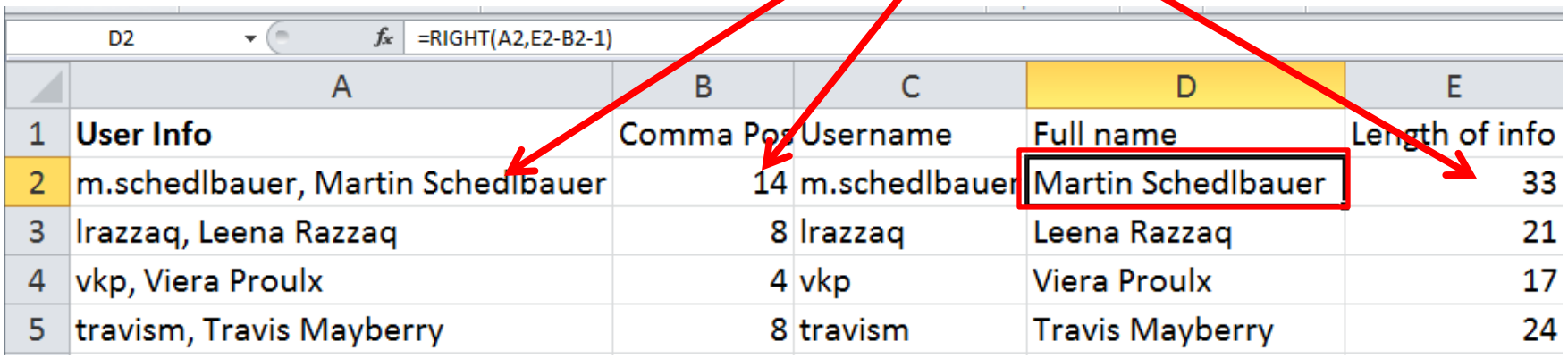
    FIND(first delimiter,cell) + LEN(first delimiter)
  - End Position = FIND(second delimiter, cell)-1
  - Number of characters =

    End position – Start Position + 1

Splitting the Text

- We could also use the MID function …

=MID(A2, B2+2, E2-(B2+2)-1)



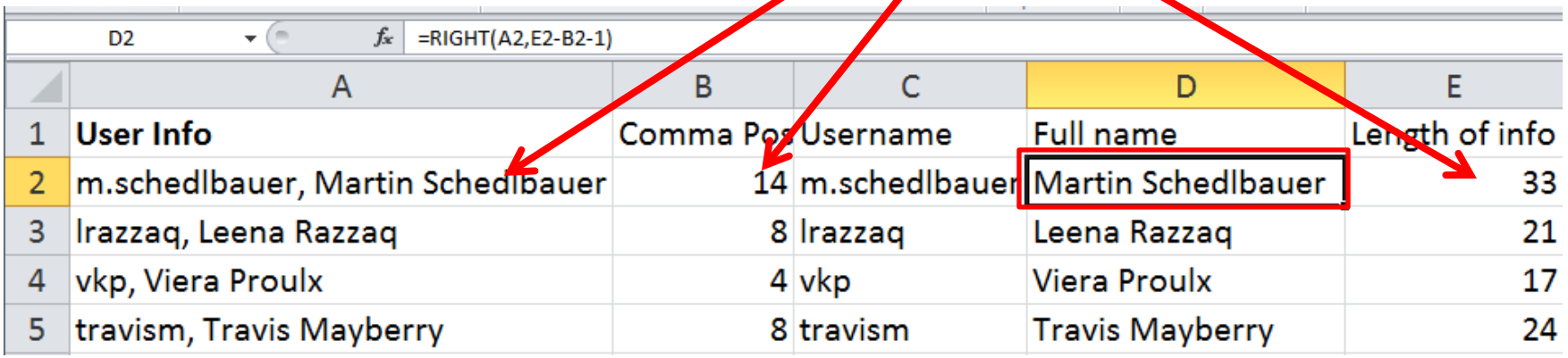
D2 =RIGHT(A2,E2-B2-1)

	A	B	C	D	E
1	**User Info**	Comma Pos	Username	Full name	Length of info
2	m.schedlbauer, Martin Schedlbauer	14	m.schedlbauer	Martin Schedlbauer	33
3	lrazzaq, Leena Razzaq	8	lrazzaq	Leena Razzaq	21
4	vkp, Viera Proulx	4	vkp	Viera Proulx	17
5	travism, Travis Mayberry	8	travism	Travis Mayberry	24

Splitting the Text

- We could also use the MID function …

=MID(A2, B2+2, E2 - B2 + 1)

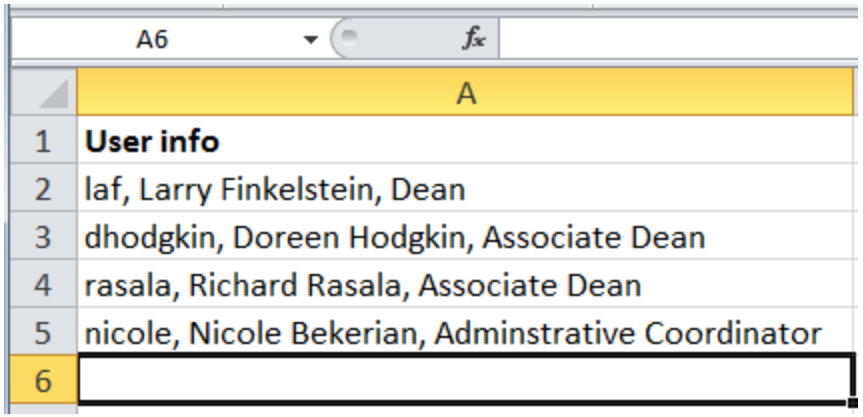


D2 =RIGHT(A2,E2-B2-1)

	A	B	C	D	E
1	**User Info**	Comma Pos	Username	Full name	Length of info
2	m.schedlbauer, Martin Schedlbauer	14	m.schedlbauer	Martin Schedlbauer	33
3	lrazzaq, Leena Razzaq	8	lrazzaq	Leena Razzaq	21
4	vkp, Viera Proulx	4	vkp	Viera Proulx	17
5	travism, Travis Mayberry	8	travism	Travis Mayberry	24

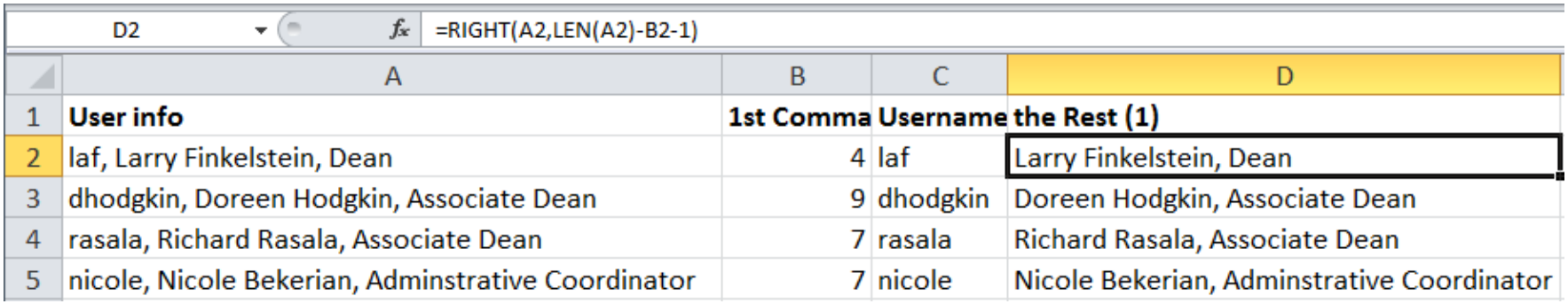
Divide and Conquer

- Divide and Conquer is a strategy for solving problems by breaking up a big problem into similar smaller problems
  - Example: suppose we are given a username, followed by a comma and a space, followed by a real name, followed by another comma and a space, followed by a job title.



Divide and Conquer
Split Once

- Our first step will be to split the original text into two parts
  1. A username
  2. Everything else



D2 =RIGHT(A2,LEN(A2)-B2-1)

	A	B	C	D
1	**User info**	**1st Comma**	**Username**	**the Rest (1)**
2	laf, Larry Finkelstein, Dean	4	laf	Larry Finkelstein, Dean
3	dhodgkin, Doreen Hodgkin, Associate Dean	9	dhodgkin	Doreen Hodgkin, Associate Dean
4	rasala, Richard Rasala, Associate Dean	7	rasala	Richard Rasala, Associate Dean
5	nicole, Nicole Bekerian, Adminstrative Coordinator	7	nicole	Nicole Bekerian, Adminstrative Coordinator

Divide and Conquer
Split Again

- Repeat the splitting process by splitting the remainder into the full name and the job title

F2 =LEFT(D2,E2-1)

	A	B	C	D	E	F	
1	**User info**	**1st Comma**	**Username**	**the Rest (1)**	**2nd Comma**	**Full name**	**Job title**
2	laf, Larry Finkelstein, Dean	4	laf	Larry Finkelstein, Dean	18	Larry Finkelstein	Dean
3	dhodgkin, Doreen Hodgkin, Associate Dean	9	dhodgkin	Doreen Hodgkin, Associate Dean	15	Doreen Hodgkin	Associate D
4	rasala, Richard Rasala, Associate Dean	7	rasala	Richard Rasala, Associate Dean	15	Richard Rasala	Associate D
5	nicole, Nicole Bekerian, Adminstrative Coordinator	7	nicole	Nicole Bekerian, Adminstrative Coc	16	Nicole Bekerian	Adminstrati

- Using this strategy, we could repeat the splitting process into smaller and smaller pieces until we have solved the problem.
- In the above example, we are done.

FIND Function

- **FIND** returns the position where a substring starts within a string.
- Optional Value: position to start search
- To find second comma: find a comma starting after the first comma.

FIND Function

- **FIND** returns the position where a substring starts within a string.
- Optional Value: position to start search

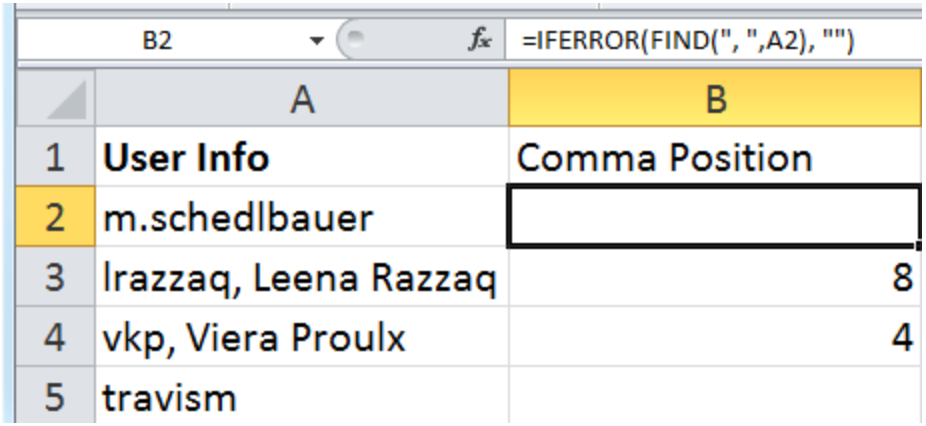
	A	B	E
1	**User info**	**1st Comma**	**2nd Comma**
2	laf, Larry Finkelstein, Dean	=FIND(", ",A2)	=FIND(", ",D2)
3	dhodgkin, Doreen Hodgkin, Associate Dean	=FIND(", ",A3)	=FIND(", ",D3)
4	rasala, Richard Rasala, Associate Dean	=FIND(", ",A4)	=FIND(", ",A4,B4+1)
5	nicole, Nicole Bekerian, Adminstrative Coordinator	=FIND(", ",A5)	=FIND(", ",A5,B5+1)

Parsing Optional Data

- Sometimes we need to split some text into parts, but one of the parts may be missing.
- A reasonable first step is to determine whether or not the data is present.

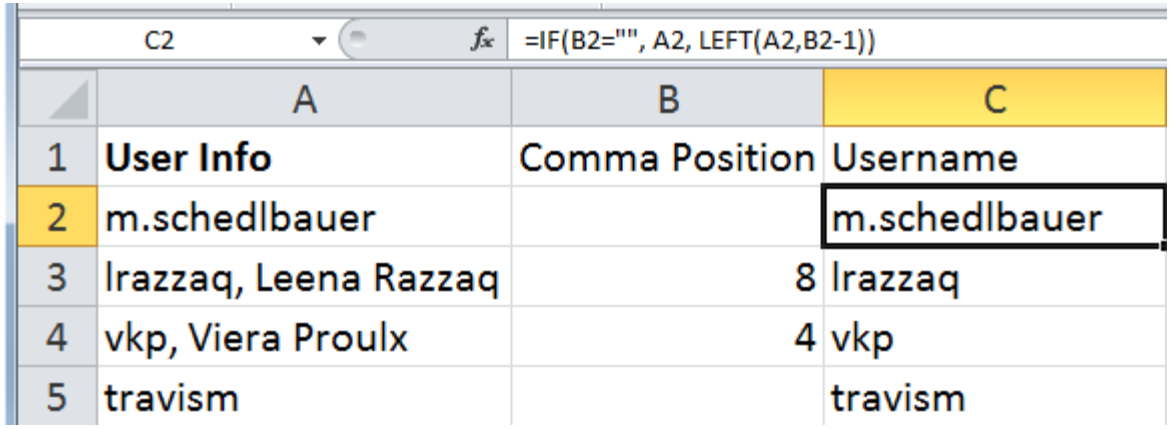
Parsing Optional Data Example

- Suppose we are given a list of usernames optionally followed by commas and a full name
- Use IFERROR and FIND to see if there is a comma and return the position if so.



Parsing Optional Data Example

- Now use an IF statement to extract the username



	A	B	C
1	**User Info**	Comma Position	Username
2	m.schedlbauer		m.schedlbauer
3	lrazzaq, Leena Razzaq	8	lrazzaq
4	vkp, Viera Proulx	4	vkp
5	travism		travism

Parsing Text

- To extract parts of a text value (parsing) requires thoughtful analysis and often a divide-and-conquer approach.

Strategy

- You need think about your strategy:
  - How do I detect where the first name starts?
  - Are there some delimiters?
  - What is the delimiter?
  - Does it always work?
  - Is there always a first or last name?
- Break the problem into several problems and create auxiliary or helper columns.

HIDDEN COLUMNS

- Solving complex parsing problems often requires the use of intermediate values:
  - Solve the problem in pieces, don't do it all in a single formula
- So, place intermediate values into temporary columns and then hide the column to make the model less confusing to read.

Let's Put This Together…

- Let's see if we can parse the text into its name and terms components…

	A	B	C
1	**Customer Information**	**Customer**	**Terms**
2	Sean White (Net30)		
3	Tim Connolly (Net10)		
4	Buck & Associates (Net30)		
5	LaSalle Construction		

- Before starting with formulas, think about your strategy.
  - How can you recognize the beginning and end of the name component?
  - How about the beginning and end of the terms component?
  - Do we need intermediate values?

COUNTA Function

- We have already seen **COUNT** as a way to count the number of cells in a range.
- However, **COUNT** only counts cells that contain numbers.
  - What about text?
- To count the number of cells that contain some value (either text or number), use **COUNTA**.

COUNTBLANK Function

- As an alternative to **COUNTA**, there is **COUNTBLANK**.
- This function counts the number of cells in a range that do not contain any value (either text or number).