Introduction to Data Analytics

Understanding Data

Tom Crawford
viznetwork.com
@viznetwork @thcrawford

University of Michigan

Introductions

About Me

About Me

Beaumont®

Valassis

root®
people.strategy.results®

masie Center

vizthink

Moveable Bytes

The IRON YARD

thinkorswim by TD Ameritrade

About You

- What's your first name?
- What's your department?
- What kind of data do you gather?
- Where do you use/want to use data visualization?
- What do you hope to get out of this class?

Today

- ✓ Introductions & Agenda
- Types of Data
- Preparing Data
- Open Data
- Practical Tips for Preparing Data
  - Strings & Numbers
  - Outliers
  - Missing Data
  - "Bad" Data

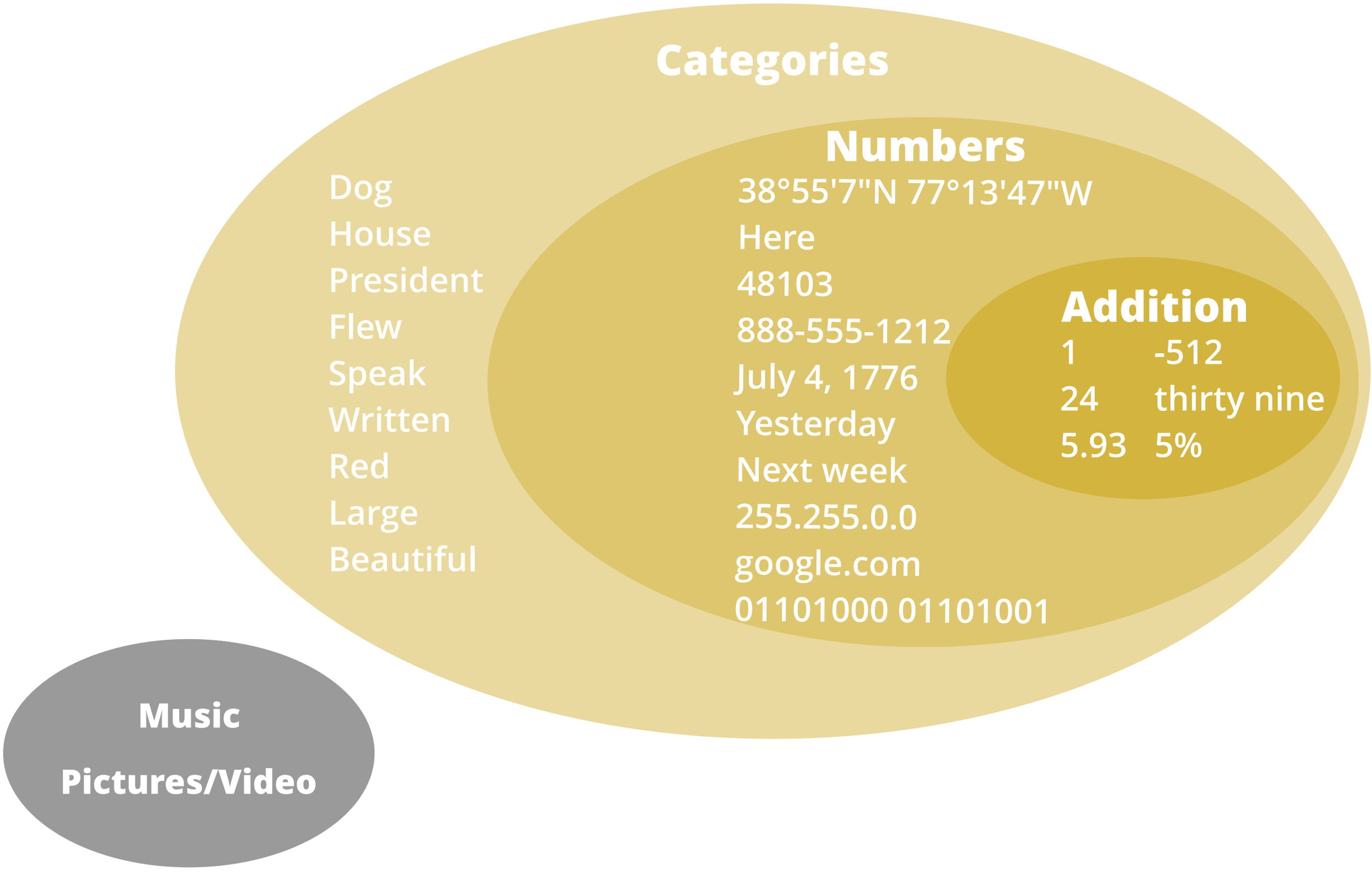
https://github.com/thcrawford/UnderstandingData

Types of Data

Where are we?

Who am I?

Types of Data



Preparing Data

1st Principle of Consistency

One Concept

For All Rows

Mary	Smith	123 Main St	Arlington
Legal Seafood		501 S Clark St	Arlington
Kite Flying	Singing	Sewing	Bowling
John	Steve	Mary	Sue

Mary	Smith	123 Main St	Arlington
Joan	Crawford	321 Oak Ave	New York City
Steve	Smith	4561 Liberty St	Ann Arbor
Mary Sue	Jones	423 State St	Lexington

2nd Principle of Consistency

One Data Type

Per Column

Mary	Smith	123 Main St	Arlington
		VA	(202) 555-9191
Steve	Smith	4561 Liberty St	Ann Arbor
		MI	(734) 555-1616

Mary	Smith	123 Main St	Arlington
Joan	Crawford	321 Oak Ave	New York City
Steve	Smith	4561 Liberty St	Ann Arbor
Mary Sue	Jones	423 State St	Lexington

Exercise

Colleague Holiday Lunch Survey

The team has decided to go out for a holiday lunch. Too many choices is a problem, so you want to have them vote on just 3 options.

The data will be tracked in a spreadsheet. What is the noun (i.e. what does the row represent) and what are its adjectives (i.e. what are its columns)?

3rd Principle of Consistency

One Format

Per Data Type

Mary	Smith	VA	(202) 555-9191
Robert Thomas	Jones	Virginia	202-618-5555
Steve	Miller Sr.	Washington DC	202.451.2323
Sonya M.	Reed	DC	+1 734-555-4545

Mary	Smith	VA	202-555-9191
Robert	Jones	VA	202-618-5555
Steve	Miller	DC	202-451-2323
Sonya	Reed	DC	734-555-4545

Exercise

Name	Address	Phone	Friends	Interests
Tom Crawford	123 Main, Ann Arbor, MI 48104	734-555-1212	Steve, Joe, Sue	Music, Food, Wine
Sue Smith	425 East First, NYC, New York	(908) 555-4545	Mary Smith	Dancing, Sewing, Music
Mr. Joe Williams	London, England	45 00 5551212		Camping, Hunting
Jonathan Jones	Europe	Unknown	To be determined	N/A

First Name	Last Name	Title	Street	City	State	Postal Code	Country	Phone
Tom	Crawford		123 Main	Ann Arbor	MI	48104	United States	734-555-1212
Sue	Smith		425 East First	New York	NY		United States	908-555-4545
Joe	Williams	Mr.		London			England	45 00 5551212
Jon	Jones							

4th Principle of Consistency

Normalize

To the Lowest Sensible Level

Normalization

Name	Address	Phone	Friends	Interests
Tom Crawford	123 Main, Ann Arbor, MI 48104	734-555-1212	Steve, Joe, Sue	Music, Food, Wine
Sue Smith	425 East First, NYC, New York	(908) 555-4545	Mary Smith	Dancing, Sewing, Music
Mr. Joe Williams	London, England	45 00 5551212		Camping, Hunting
Jonathan Jones	Europe	Unknown	To be determined	N/A

First Name	Last Name	Title	Street	City	State	Postal Code	Country	Phone
Tom	Crawford		123 Main	Ann Arbor	MI	48104	United States	734-555-1212
Sue	Smith		425 East First	New York	NY		United States	908-555-4545
Joe	Williams	Mr.		London			England	45 00 5551212
Jon	Jones							

Normalization

ID	First Name	Last Name	Title	Street	City	State	Postal Code
1	Tom	Crawford		123 Main	Ann Arbor	MI	48104
2	Sue	Smith		425 East First	New York	NY	
3	Joe	Williams	Mr.		London		
4	Jon	Jones					

PersonID	ActivityID
1	1
1	2
1	3
2	4
2	5
2	1
3	6
3	7

ID	Activity
1	Music
2	Food
3	Wine
4	Dancing
5	Sewing
6	Camping
7	Hunting

Exercise

Colleague Holiday Lunch Survey Redux

The team didn't like your 3 choices. They want to make their own (unlimited) recommendations and have people choose as many as they like.

The data will be tracked in a spreadsheet. What is/are the noun(s) and what are its/their adjectives?

Exercise

Weight Loss Study

One of the researchers at the University of Michigan has come up with an incredible new weight-loss program. To confirm the efficacy, we will need to track people in the program including intake, exercise, and, of course, results. Due to the sensitive nature of the results, only the research coordinator can see the names and other identifying characteristics. The researchers and statisticians should not have access to identifying information.

The data will be tracked in a spreadsheet. What is/are the noun(s) and what are its/their adjectives?

Open Data

Sample Sources

DATA.GOV

United States™ Census Bureau

The Opportunity Project

Michigan.gov

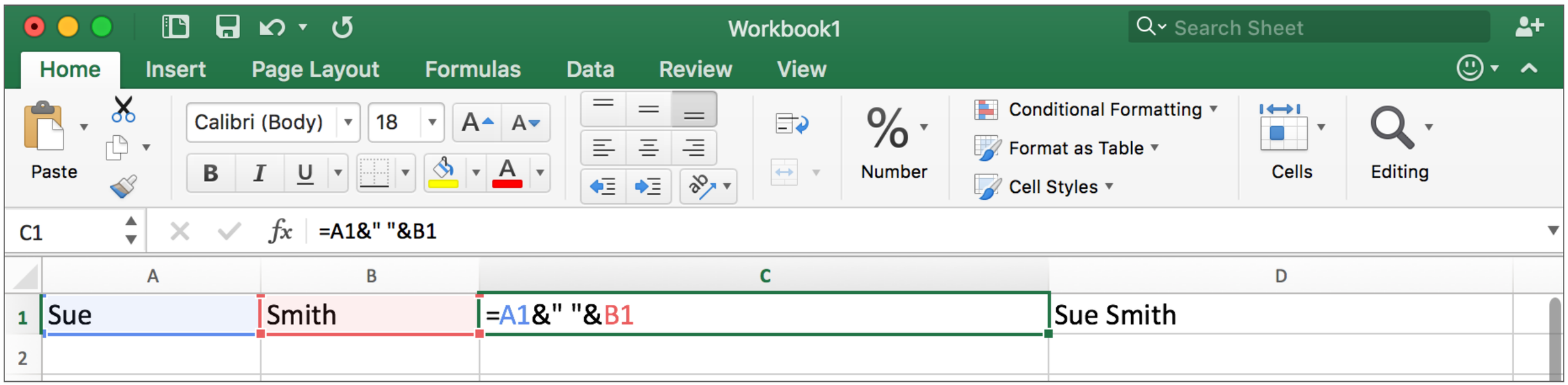
DATA DRIVEN DETROIT

InformedCity

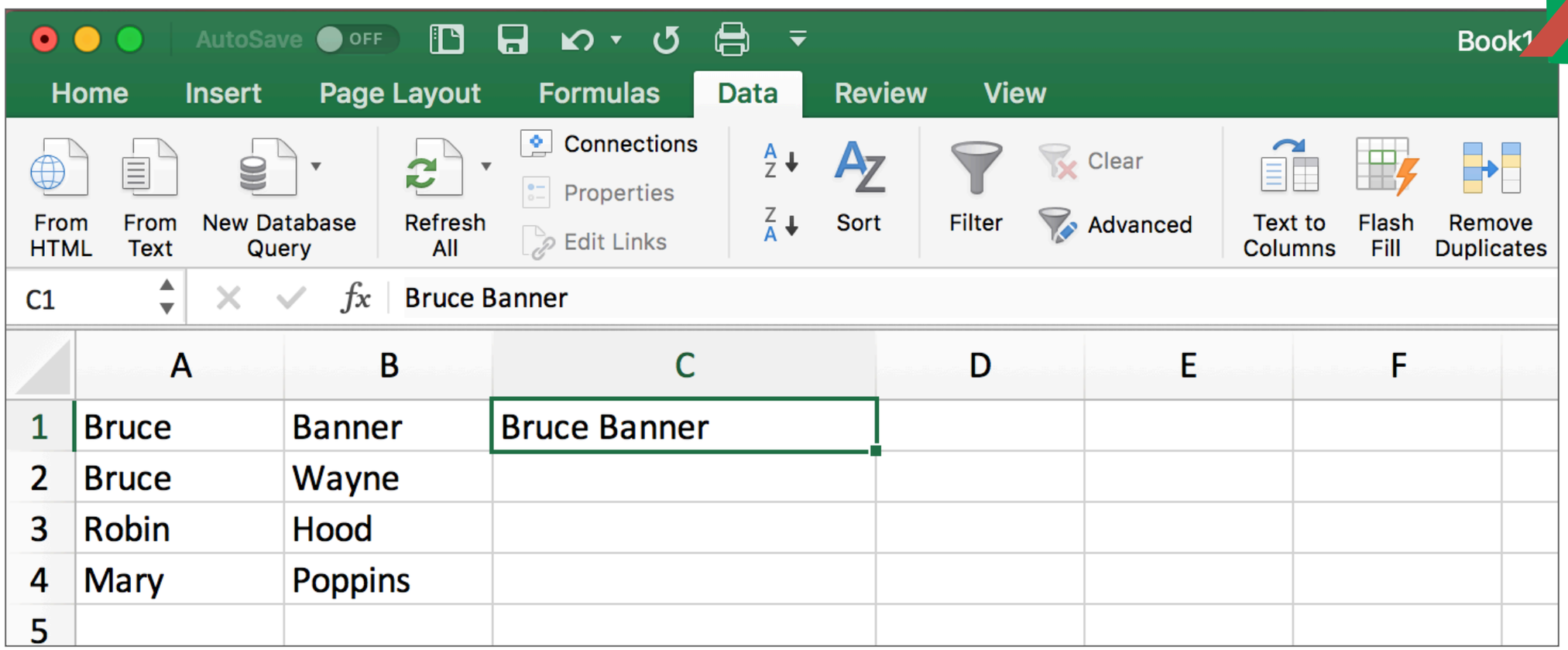
Tips for Excel (and Google Sheets)

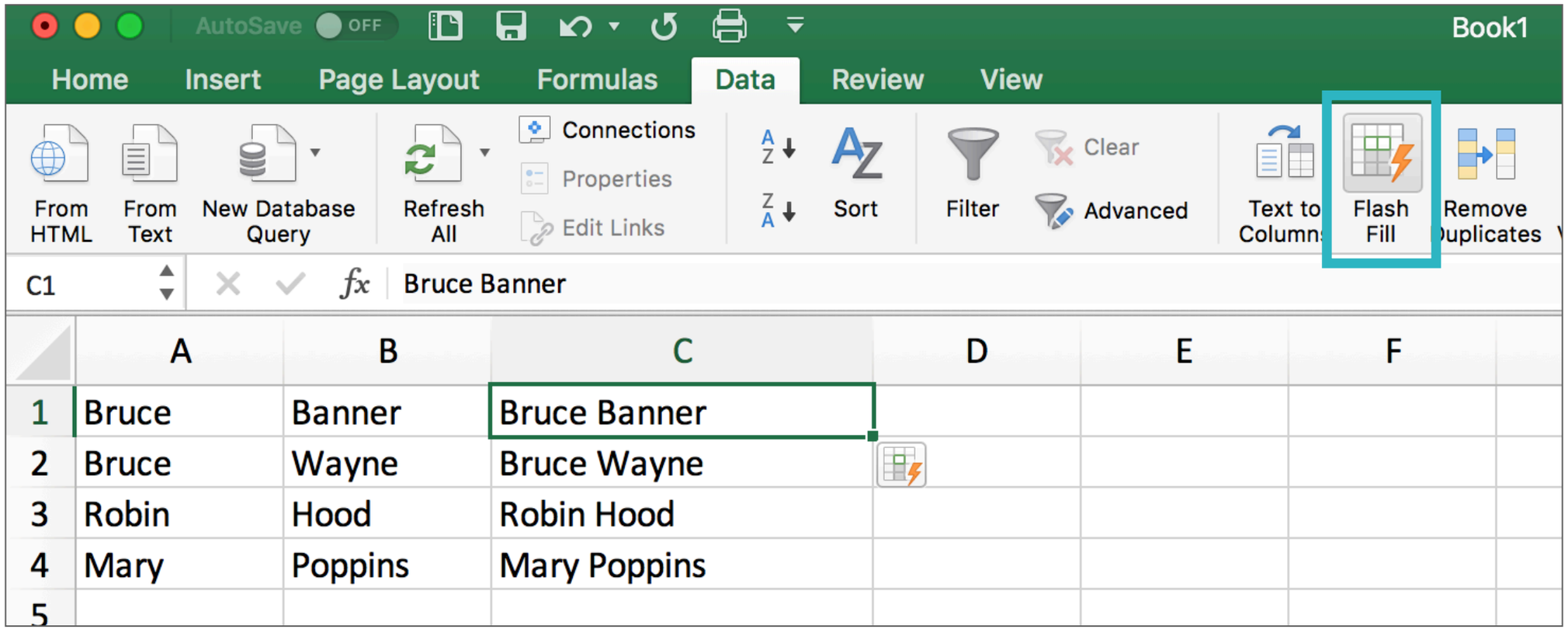
Strings

Joining Strings

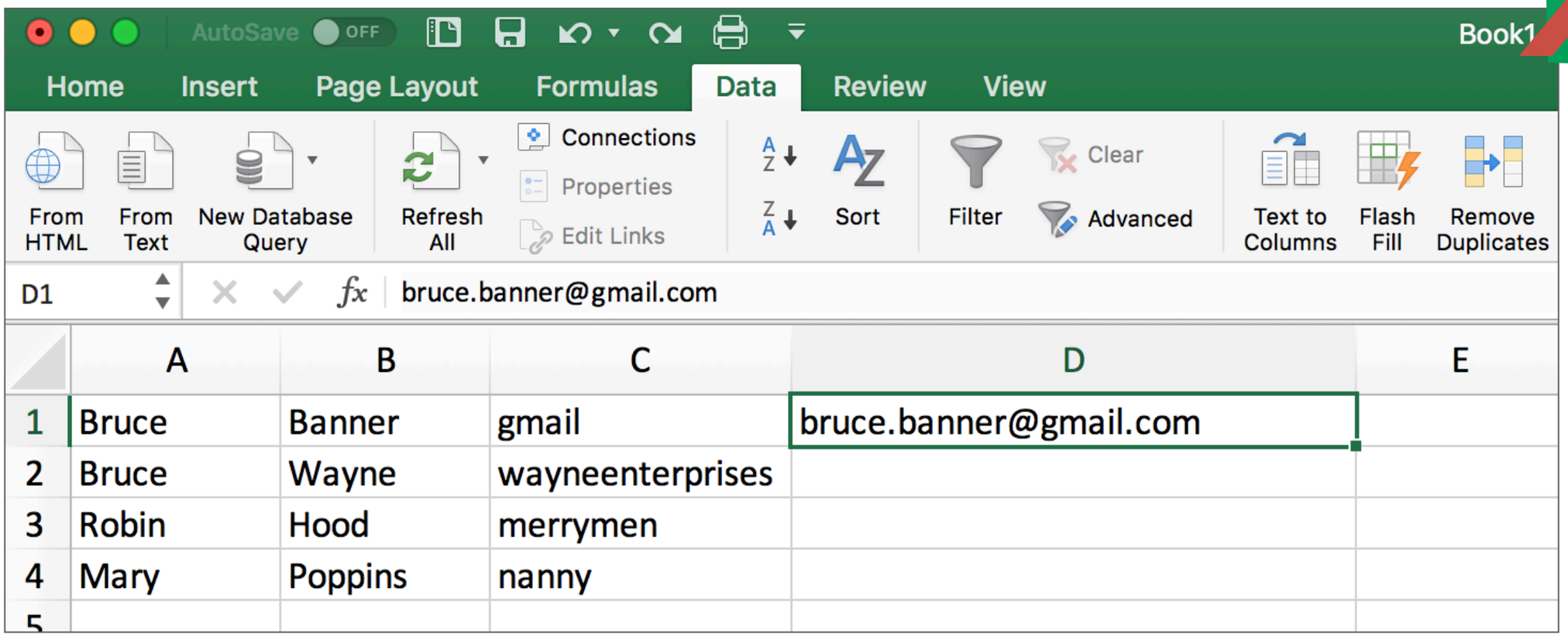


Joining Strings

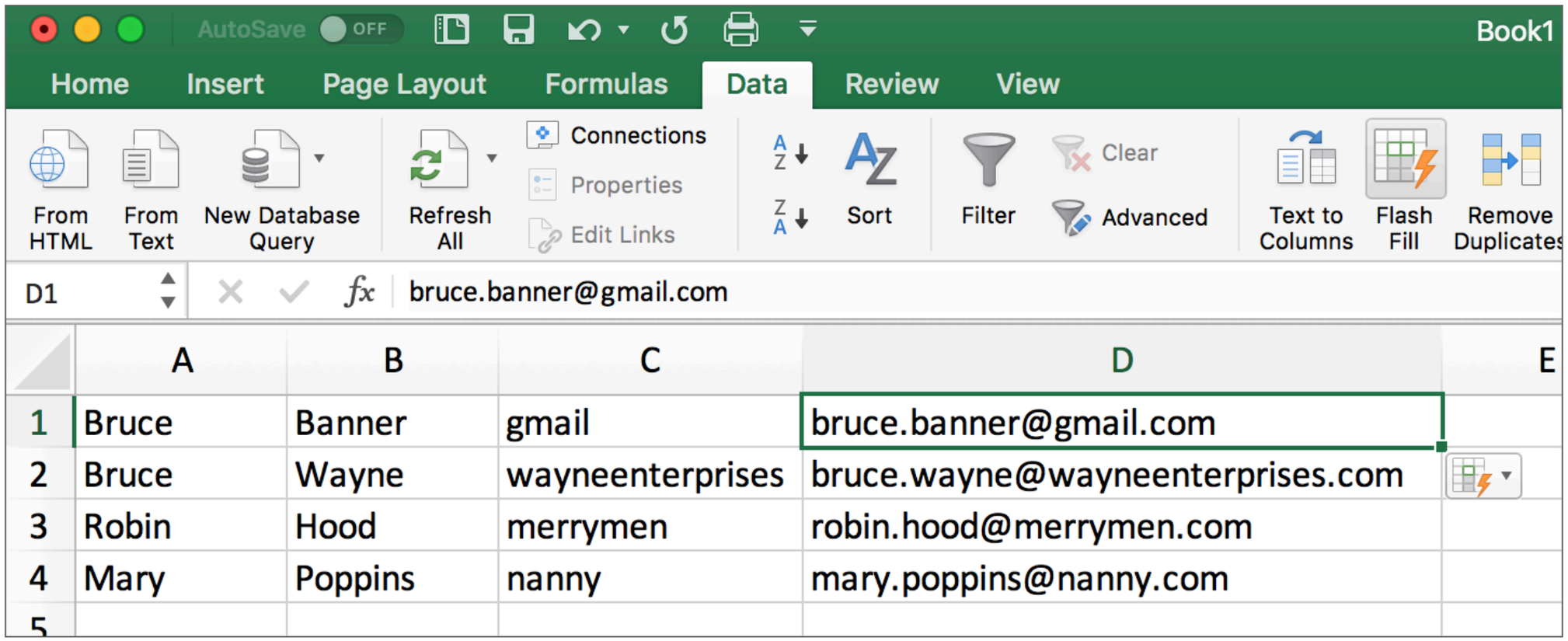




Joining Strings

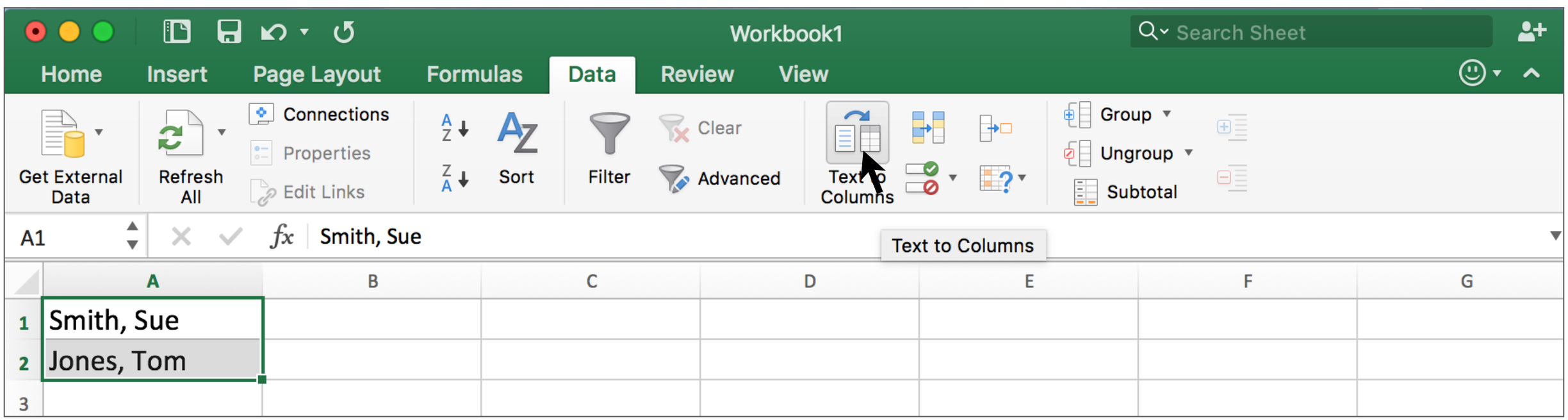


	A	B	C	D	E
1	Bruce	Banner	gmail	bruce.banner@gmail.com	
2	Bruce	Wayne	wayneenterprises		
3	Robin	Hood	merrymen		
4	Mary	Poppins	nanny		
5					

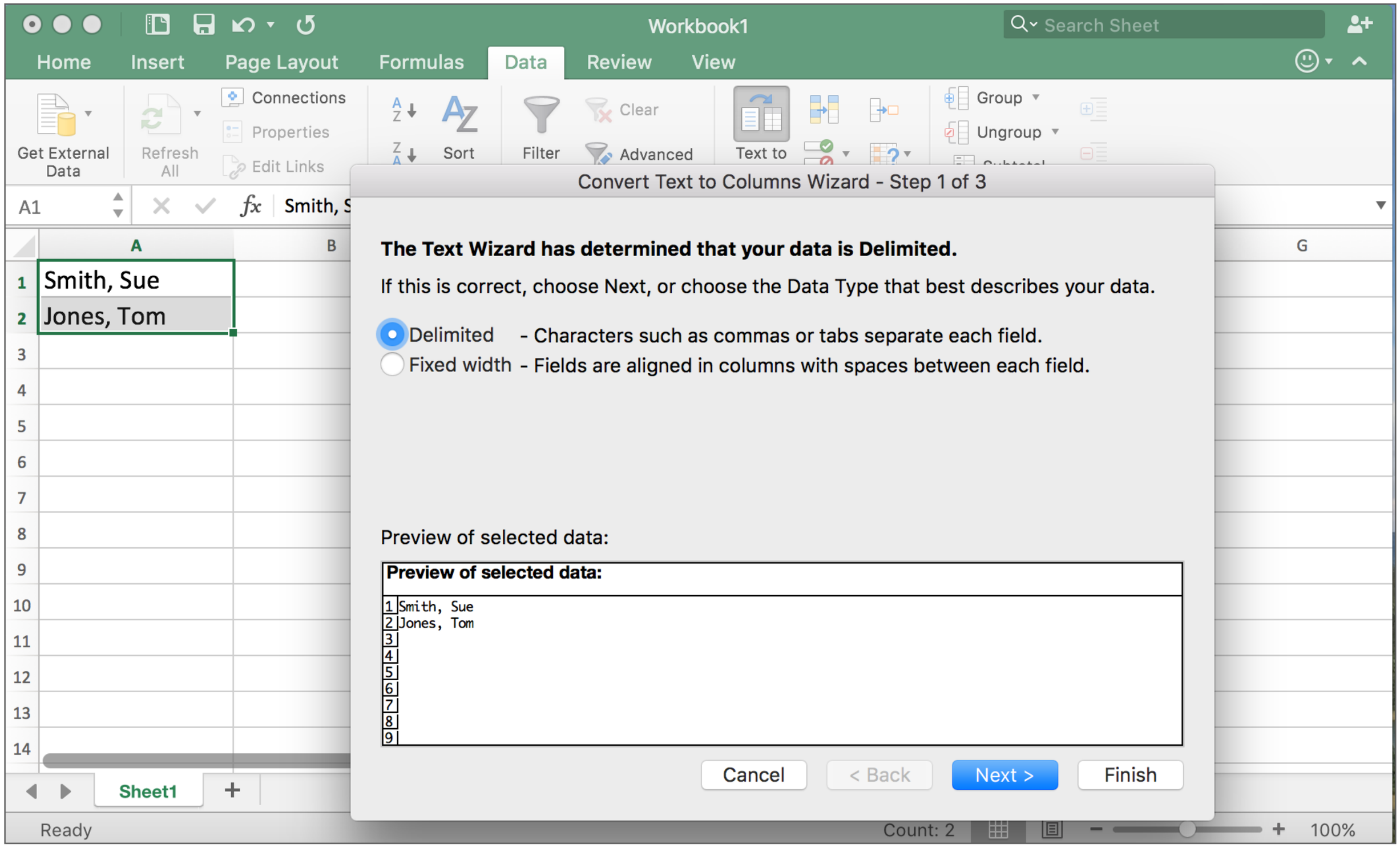


	A	B	C	D	E
1	Bruce	Banner	gmail	bruce.banner@gmail.com	
2	Bruce	Wayne	wayneenterprises	bruce.wayne@wayneenterprises.com	
3	Robin	Hood	merrymen	robin.hood@merrymen.com	
4	Mary	Poppins	nanny	mary.poppins@nanny.com	
5					

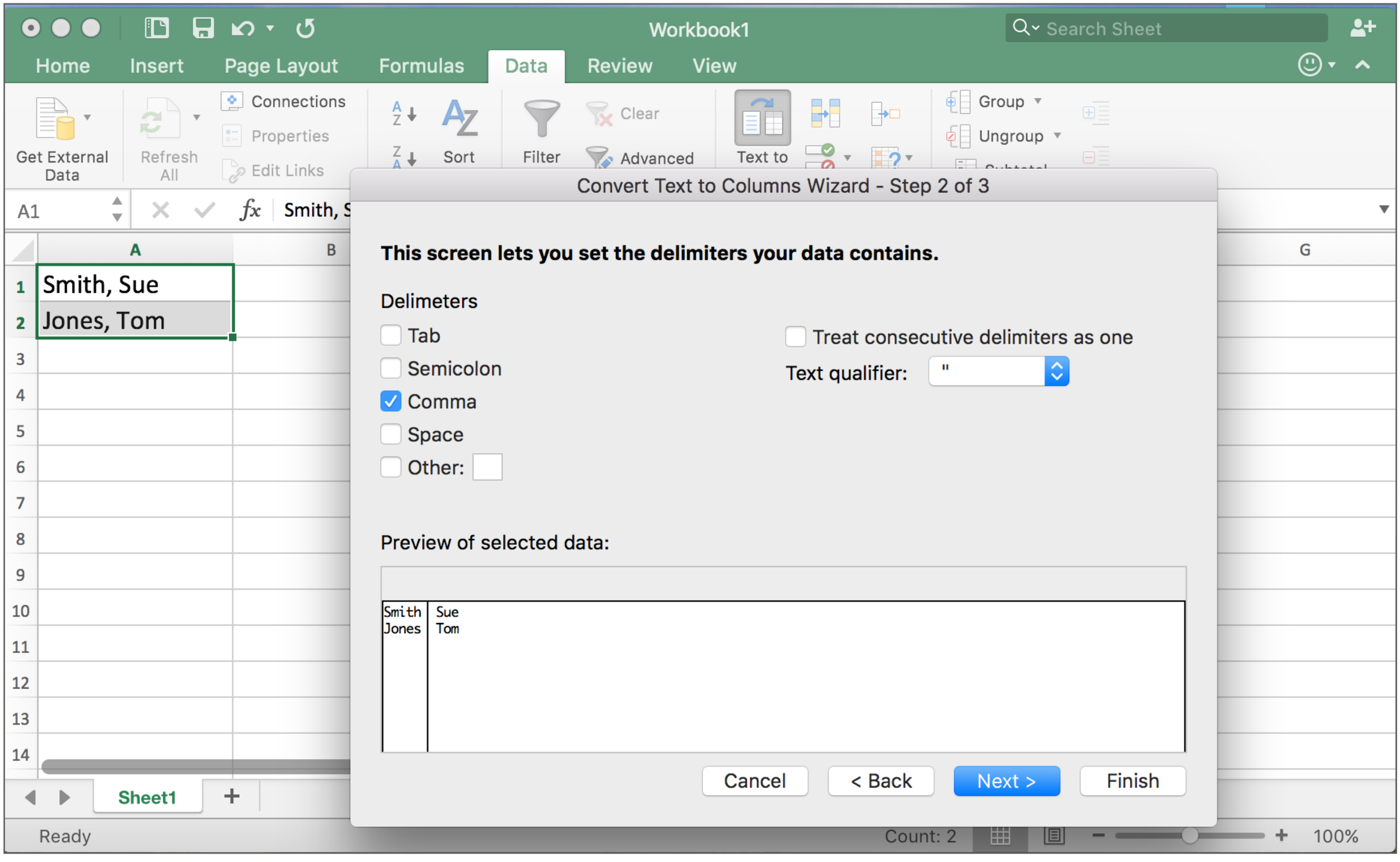
Parsing Strings



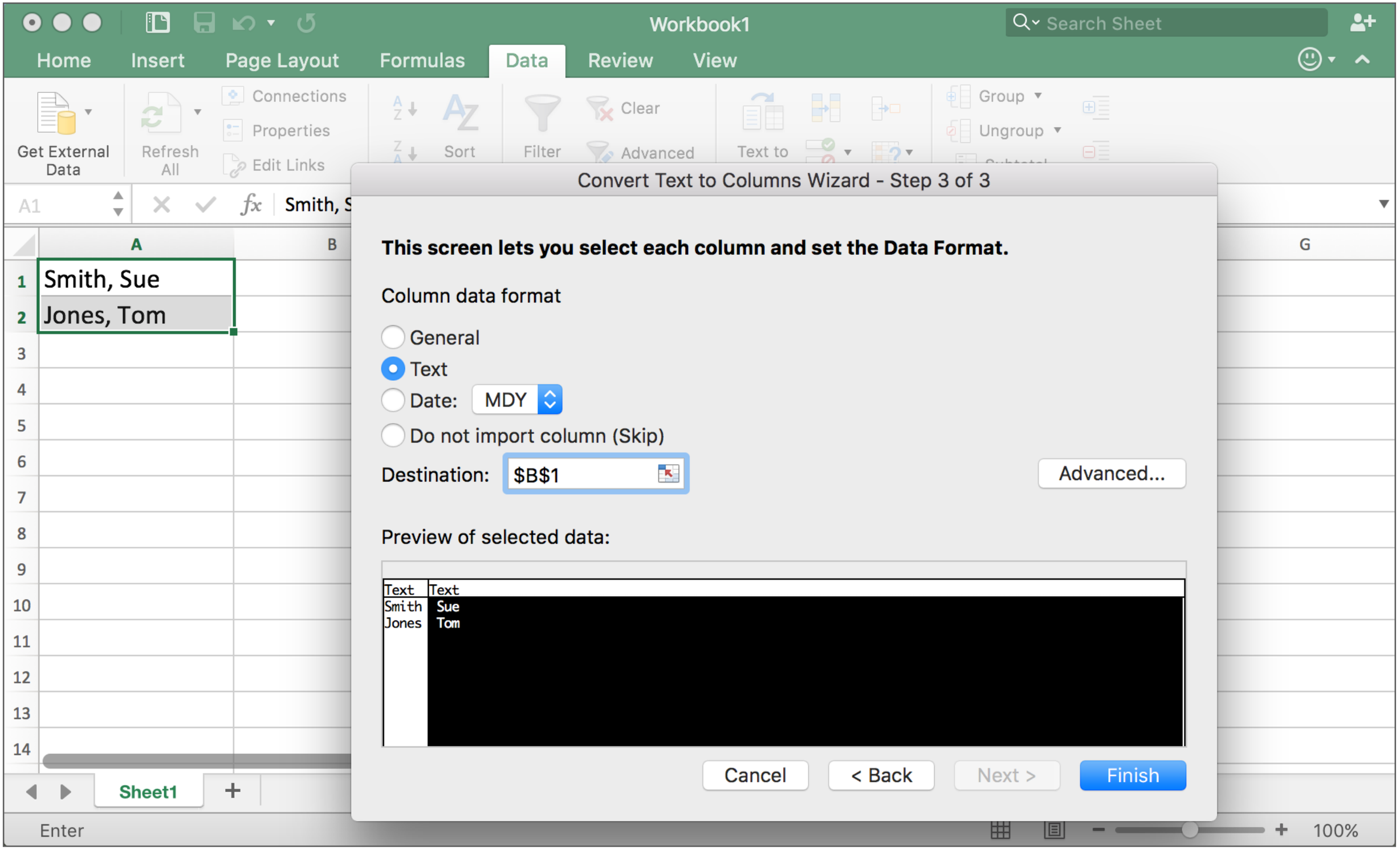
Parsing Strings



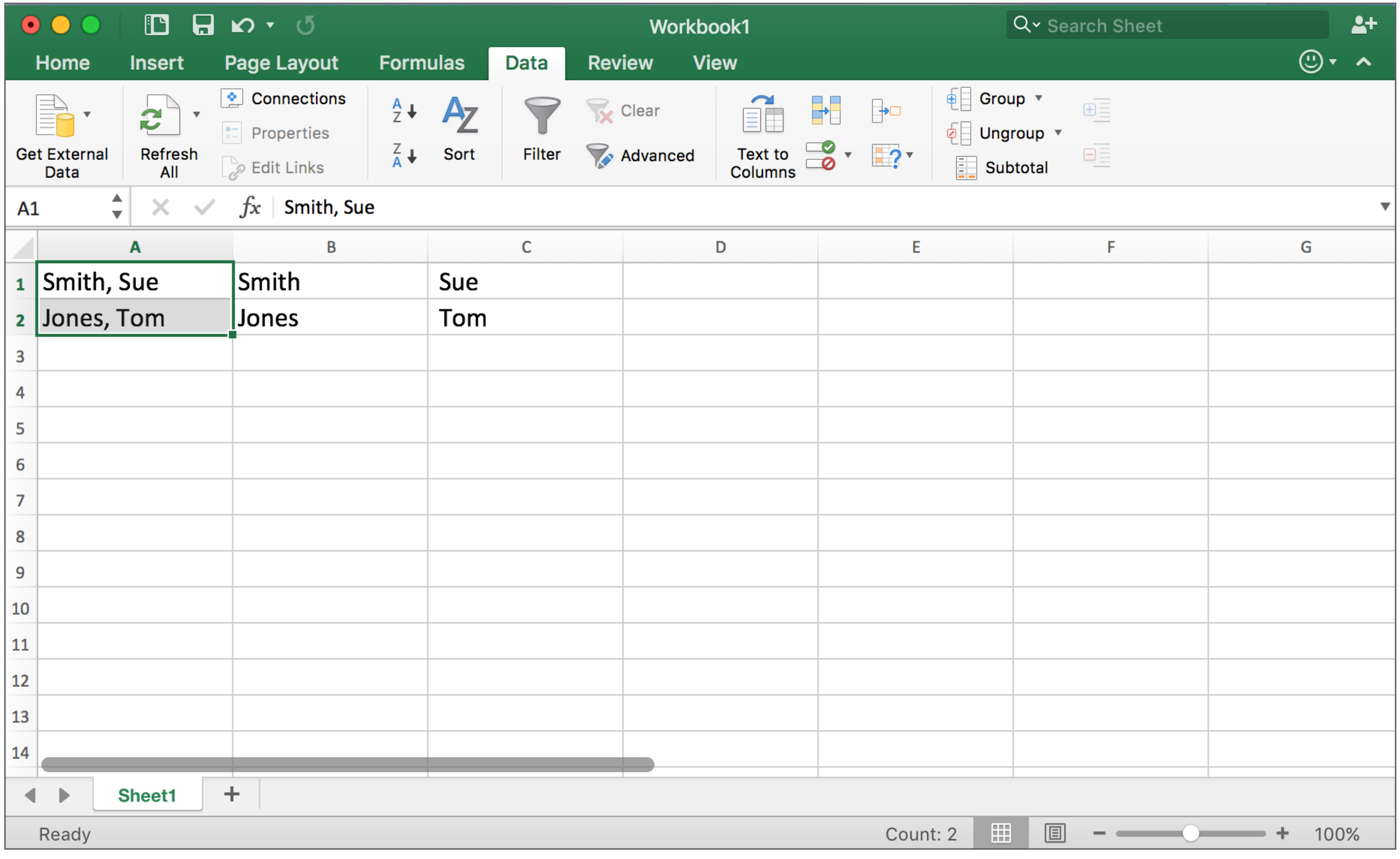
Parsing Strings



Parsing Strings



Parsing Strings



Splitting Strings

Workbook1

Search Sheet

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Calibri (Body) | 22 | Number | Conditional Formatting | Format as Table | Cell Styles | Cells | Editing

H1

	A	B	C	D	E	F	G	H	I
1	Jessica Jones	8	9	0	Jessica		Jones		
2	Matthew Michael Murdock	8	9	16	Matthew	Michael	Murdock		
3	Bruce Banner	6	7	0	Bruce		Banner		

Workbook1

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Cut | Copy | Format | Calibri (Body) | 22 | Wrap Text | General | Merge & Center

H1

	A	B	C	D
1	Jessica Jones	=SEARCH(" ",A1)	=B1 + 1	=IFERROR(SEARCH(" ",A1,C1),0)
2	Matthew Michael Murdock	=SEARCH(" ",A2)	=B2 + 1	=IFERROR(SEARCH(" ",A2,C2),0)
3	Bruce Banner	=SEARCH(" ",A3)	=B3 + 1	=IFERROR(SEARCH(" ",A3,C3),0)

Splitting Strings

Workbook1

Search Sheet

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Calibri (Body) | 22 | Number | Conditional Formatting | Format as Table | Cell Styles | Cells | Editing

H1

	A	B	C	D	E	F	G	H	I
1	Jessica Jones	8	9	0	Jessica		Jones		
2	Matthew Michael Murdock	8	9	16	Matthew	Michael	Murdock		
3	Bruce Banner	6	7	0	Bruce		Banner		

Workbook1

Search Sheet

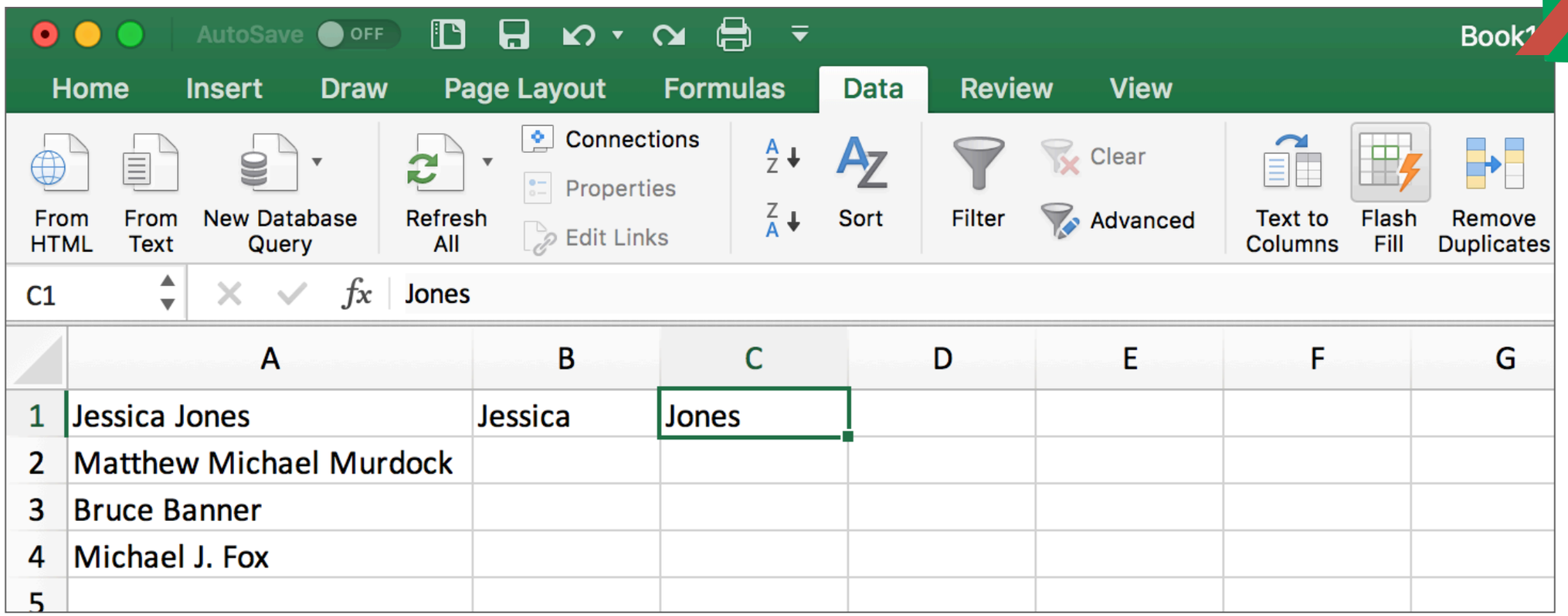
View

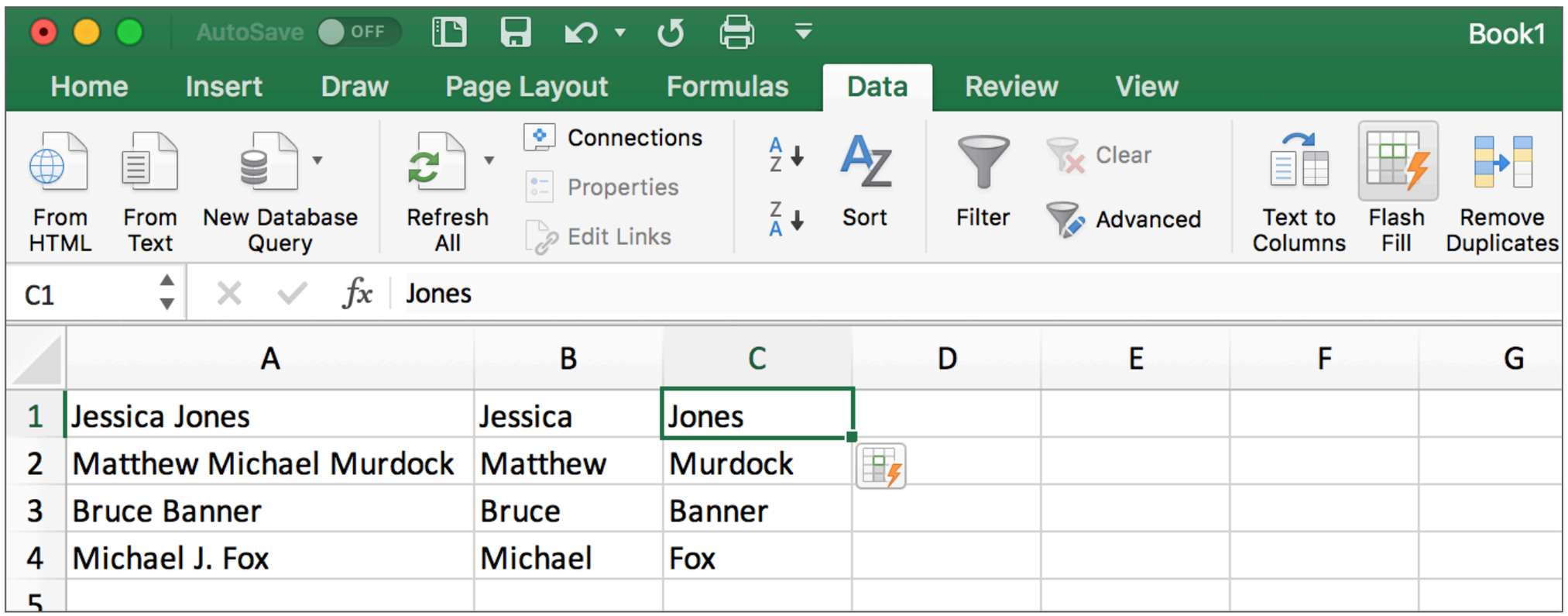
Wrap Text | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | AutoSum | Fill | Clear | Sort & Filter

Merge & Center

E	F	G
=LEFT(A1,B1)	=IF(D1=0,"",MID(A1,B1,D1-B1))	=IF(D1=0,RIGHT(A1,LEN(A1)-B1),RIGHT(A1,LEN(A1)-D1))
=LEFT(A2,B2)	=IF(D2=0,"",MID(A2,B2,D2-B2))	=IF(D2=0,RIGHT(A2,LEN(A2)-B2),RIGHT(A2,LEN(A2)-D2))
=LEFT(A3,B3)	=IF(D3=0,"",MID(A3,B3,D3-B3))	=IF(D3=0,RIGHT(A3,LEN(A3)-B3),RIGHT(A3,LEN(A3)-D3))

Splitting Strings



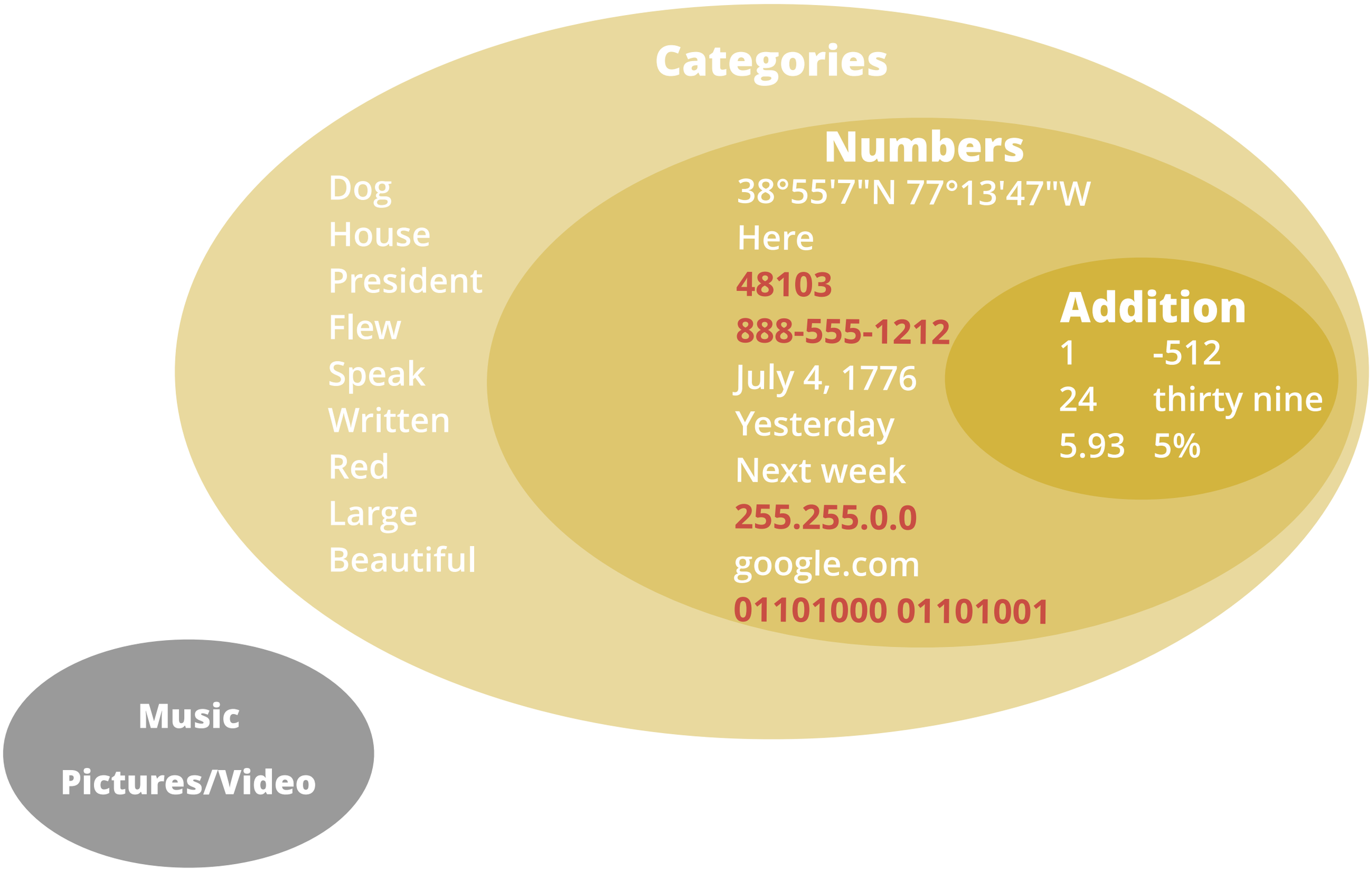


Important Spreadsheet Functions

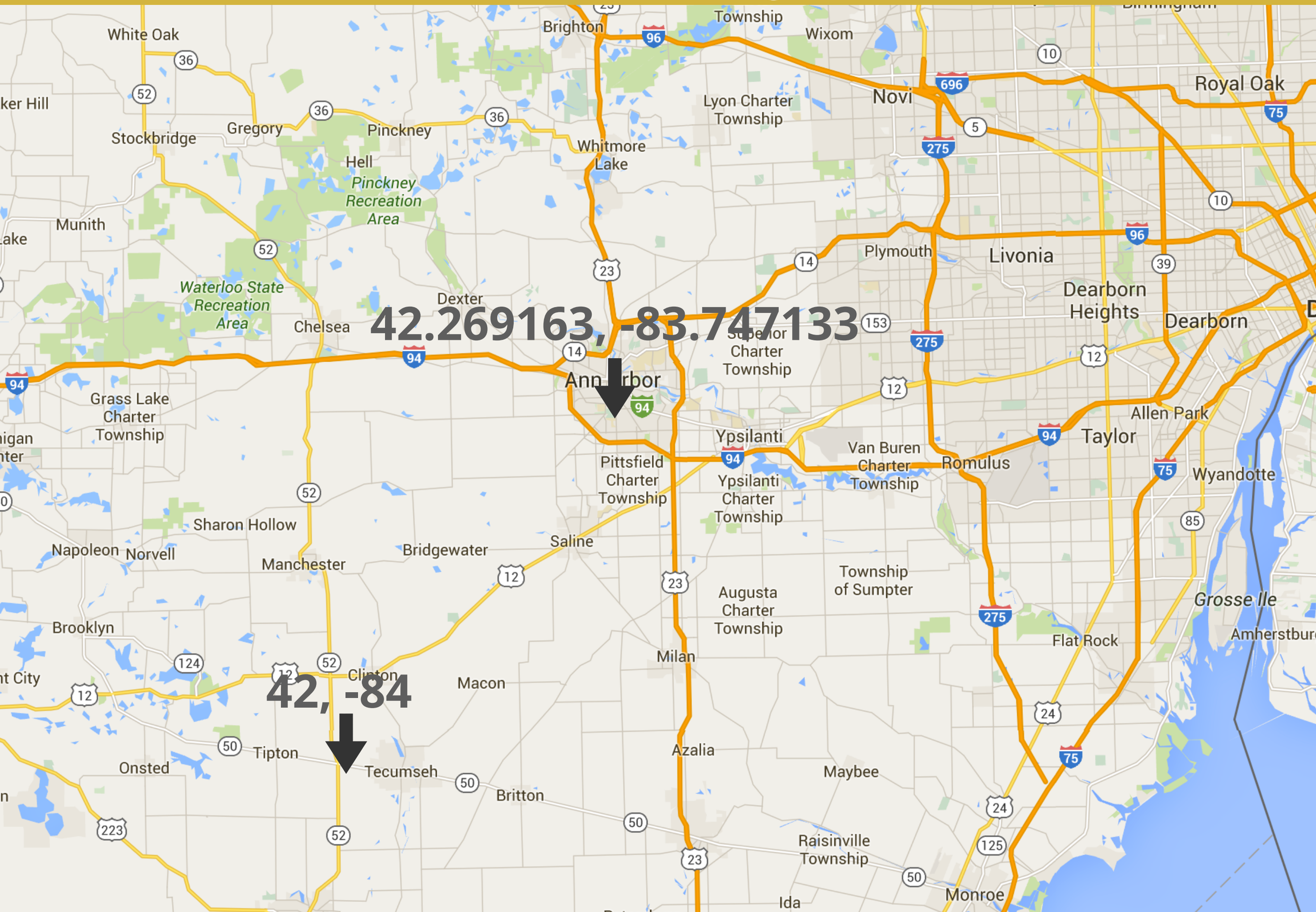
- & …to join strings
- Data > Text to Columns …to separate consistent strings
- SEARCH() …to find the location of a string in another string
- LEFT() …to get the beginning of a string
- RIGHT() …to get the end of a string
- MID() …to get the middle of a string
- LEN() …to get the length of a string
- IFERROR() …to suppress an error with another result

Numbers

Numbers That Aren't Numbers



Rounding


42.269163, -83.747133
42, -84
White Oak
Brighton
Township
Wixom
Novi
Royal Oak
Lyon Charter Township
Stockbridge
Gregory
Pinckney
Whitmore Lake
Hell
Pinckney Recreation Area
Munith
Waterloo State Recreation Area
Dexter
Chelsea
Plymouth
Livonia
Dearborn Heights
Dearborn
Superior Charter Township
Ann Arbor
Grass Lake Charter Township
Ypsilanti
Pittsfield Charter Township
Ypsilanti Charter Township
Van Buren Charter Township
Romulus
Taylor
Allen Park
Wyandotte
Sharon Hollow
Napoleon
Norvell
Manchester
Bridgewater
Saline
Township of Sumpter
Augusta Charter Township
Grosse Ile
Brooklyn
Clinton
Macon
Milan
Flat Rock
Onsted
Tipton
Tecumseh
Britton
Azalia
Maybee
Raisinville Township
Monroe
Ida

Identify Outliers

Outliers

Causes:

- Bad measurement
- Incorrect data entry
- Fraud
- Natural variation
- Flawed theory

Solutions:

- Retain
- Exclude
- Explore other models
- Re-sample
- Be aware!

Handle Missing Data

Infer Missing Data

Existing Data: 48103
Infer: Ann Arbor, Michigan, Eastern Time, United States, North America, Northern Hemisphere…

Existing Data: United States
Infer: North America, Northern Hemisphere…

Existing Data: Blue Eyes
Infer: Nothing

Handle Missing Data

- Avoid using N/A, Not Available, or any other string
- Do not enter 0 (zero) unless the value is actually zero
- Preferably leave unknown fields blank
- Be aware that some statistics programs handle blanks, nulls, spaces, and zeros differently, and know how your system will handle them

	A	B
1	1	"1"
2	0	"0"
3		" "
4		(empty)
5	1	
6	2	
7	3	

	A	B
1	1	"1"
2	0	"0"
3		" "
4		(empty)
5	=SUM(A1:A4)	
6	=COUNT(A1:A4)	
7	=COUNTA(A1:A4)	

Test for “Bad” Data

Test for “Bad” Data: Parsing

Delimiters in the text

"1","Onion, Large", "Diced"

Quotation marks in the text

"6","Dice into 1/3" pieces"

Slashes interpreted as escapes

"Good/New"

Ampersands (&) in data
Line Feeds & Carriage Returns
Multiple dash types

Test for “Bad” Data: Characters

ASCII vs UTF8 vs UTF16 vs other encodings
Capitalization
Ellipses vs Dots

. . . . . .

“Smart quotes” vs. "quotes"…

“x” "x"

Diacritics…

cafe vs. café

Test for "Bad" Data: Multiple Formats

Name formats
Websites in email fields & vice versa
Rounding numbers
Money formats
Money conversions
Date formats (including "yesterday")
Phone number formats
  (Twillio API)
Address formats
  (USPS API)
Zip codes treated as numbers (i.e. Maine & Canada)

Exercise

BB-8 Infection Tracking

In early 2015, there were early indications of a new communicable infection that seems to be coming from somewhere in northern California. It seems to be being spread by marketers. We need to interview individuals that have contracted the infection to find out where they've been and who they've been in contact with. In the end, we will need to use the data to create a network/tree diagram to help identify Patient 0.

The data will be tracked in a spreadsheet. What is/are the noun(s) and what are its/their adjectives?

Displaying Data