



# Introduction to **Data Analytics**

## Understanding Data

University of Michigan

Tom Crawford  
[viznetwork.com](http://viznetwork.com)  
@viznetwork @thcrawford

# Introductions

# About Me



# About Me



Beaumont<sup>®</sup>

Valassis

root.  
people.strategy.results.<sup>®</sup>

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

## Categories

Dog  
House  
President  
Flew  
Speak  
Written  
Red  
Large  
Beautiful

**Music**  
**Pictures/Video**

## Numbers

38°55'7"N 77°13'47"W  
Here  
48103  
888-555-1212  
July 4, 1776  
Yesterday  
Next week  
255.255.0.0  
google.com  
01101000 01101001

## Addition

1	-512
24	thirty nine
5.93	5%

# 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	Xewing	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
		X	(202) 555-9191
Steve	Smith	561 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				Postal				
Name	Last Name	Title	Street	City	State	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

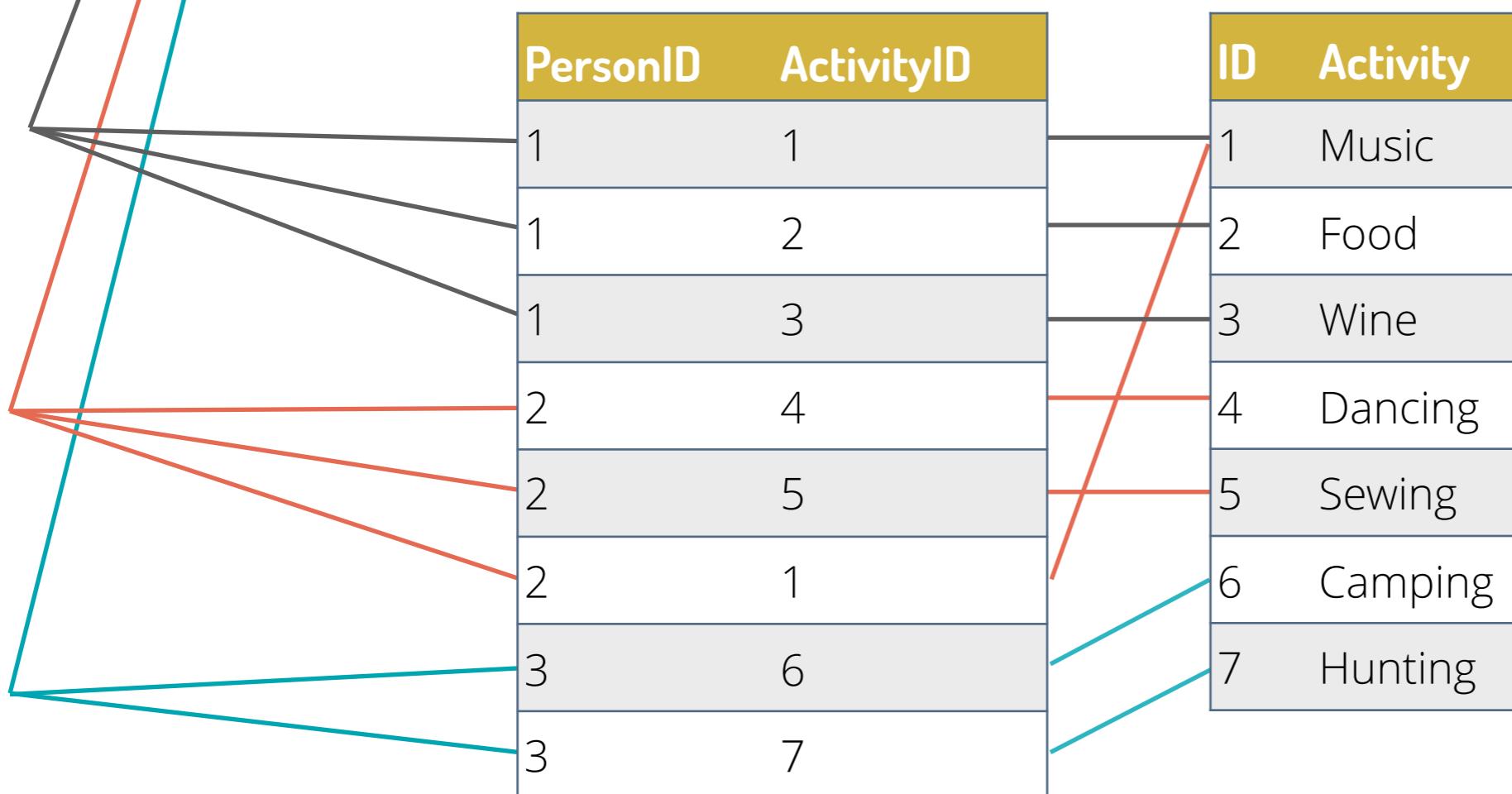
# Normalization

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					



# 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

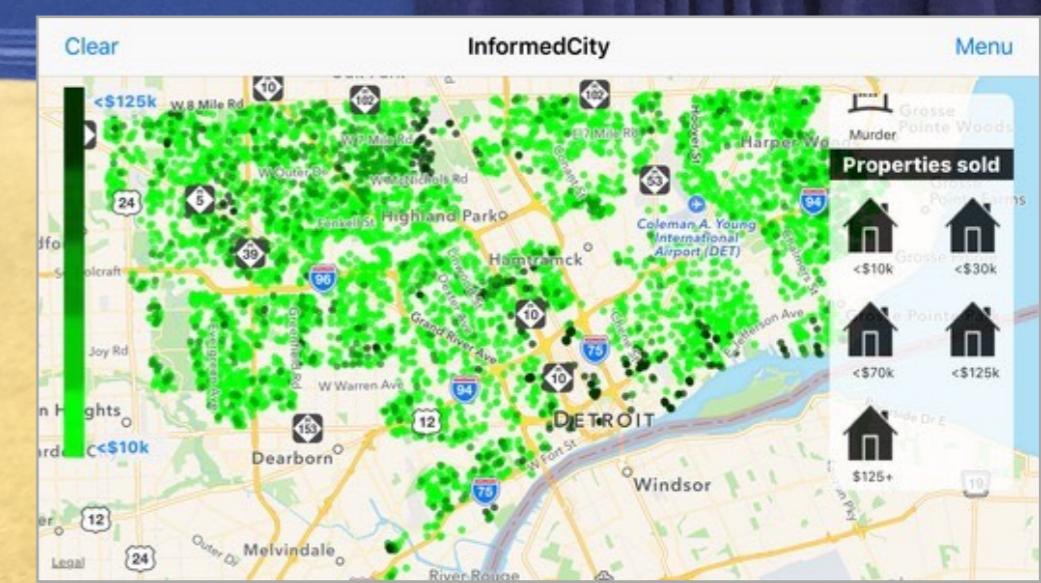
# Sample Sources



The Opportunity Project



# InformedCity



# Tips for Excel (and Google Sheets)

# Strings

# Joining Strings

Screenshot of Microsoft Excel showing the formula bar and a simple string concatenation example.

The formula bar shows the formula `=A1&" "&B1` entered into cell C1. The cell contains the result "Sue Smith".

The Excel ribbon is visible at the top, showing tabs like Home, Insert, Page Layout, Formulas, Data, Review, and View. The Home tab is selected.

The formula bar also displays the current font (Calibri Body), font size (18), and other editing tools.

	A	B	C	D
1	Sue	Smith	=A1&" "&B1	Sue Smith
2				

# Joining Strings

This screenshot shows a Microsoft Excel spreadsheet with the Data tab selected. A green box highlights the cell C1, which contains the text "Bruce Banner". The range C1:C5 is selected, and the formula bar shows "Bruce Banner" in the fx field. The Data ribbon tab has a red X mark over it, indicating an error or a specific point of focus.

	A	B	C	D	E	F
1	Bruce	Banner	Bruce Banner			
2	Bruce	Wayne				
3	Robin	Hood				
4	Mary	Poppins				
5						

This screenshot shows the same Microsoft Excel spreadsheet after using the Flash Fill feature. The Data ribbon tab now has a blue box around the "Flash Fill" icon, which is highlighted with a cyan box. The formula bar still shows "Bruce Banner" in the fx field. The Data ribbon tab is now functional and correctly displays the joined names in column C.

	A	B	C	D	E	F
1	Bruce	Banner	Bruce Banner			
2	Bruce	Wayne	Bruce Wayne			
3	Robin	Hood	Robin Hood			
4	Mary	Poppins	Mary Poppins			
5						

# Joining Strings



Screenshot of Microsoft Excel showing a simple string concatenation. The formula bar shows the formula `=bruce.banner@gmail.com`. The data table contains five rows of names and their email suffixes.

	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					

Screenshot of Microsoft Excel showing a more complex string concatenation using the Flash Fill feature. The formula bar shows the formula `=bruce.banner@gmail.com`. The data table contains five rows of names and their email suffixes, with the last row partially visible.

	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



Workbook1

Search Sheet

Home Insert Page Layout Formulas Data Review View

Get External Data Refresh All Connections Properties Edit Links

A Z Sort Filter Advanced

Text to Columns Group Ungroup Subtotal

A1 Smith, Sue

Text to Columns

	A	B	C	D	E	F	G
1	Smith, Sue						
2	Jones, Tom						
3							

# Parsing Strings

Workbook1

Search Sheet

Home Insert Page Layout Formulas Data Review View

Get External Data Refresh All Edit Links

A1 Smith, S

A B G

1 Smith, Sue  
2 Jones, Tom

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the Data Type that best describes your data.

Delimited - Characters such as commas or tabs separate each field.  
 Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

Preview of selected data:

1 Smith, Sue  
2 Jones, Tom  
3  
4  
5  
6  
7  
8  
9

Cancel < Back Next > Finish

Count: 2 100%

Sheet1

# Parsing Strings

Workbook1

Search Sheet

Home Insert Page Layout Formulas Data Review View

Get External Data Refresh All Properties Edit Links

A1 Smith, S

A B G

1 Smith, Sue  
2 Jones, Tom

This screen lets you set the delimiters your data contains.

Delimiters

Tab  
 Semicolon  
 Comma  
 Space  
 Other:

Treat consecutive delimiters as one  
Text qualifier: "

Preview of selected data:

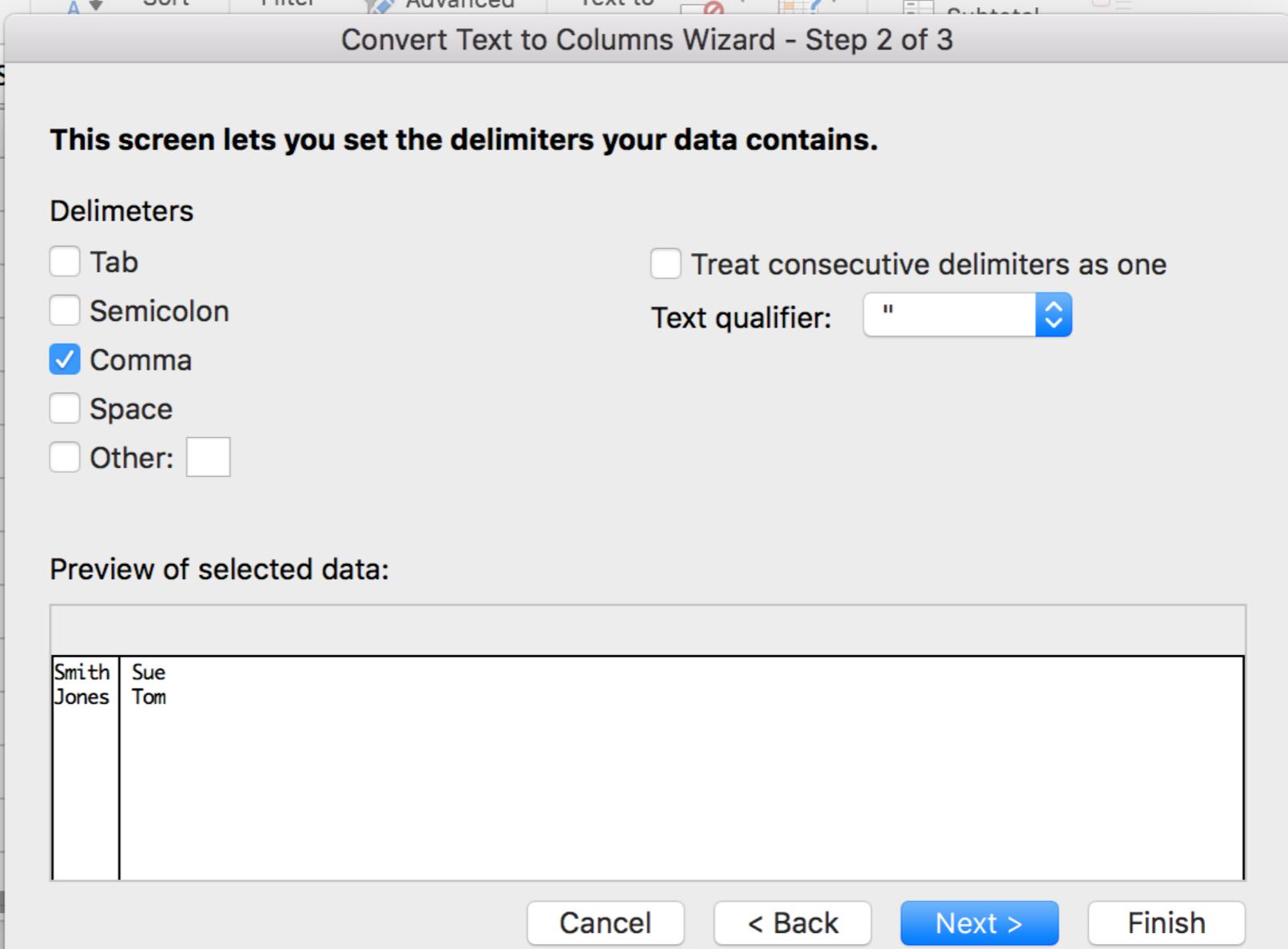
Smith	Sue
Jones	Tom

Cancel < Back Next > Finish

Count: 2 100%

Sheet1

Ready



# Parsing Strings

Workbook1

Search Sheet

Home Insert Page Layout Formulas Data Review View

Get External Data Refresh All Properties Edit Links

A1 Smith, S Jones, Tom

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

General  
 Text  
 Date: MDY  
 Do not import column (Skip)

Destination: \$B\$1 Advanced...

Preview of selected data:

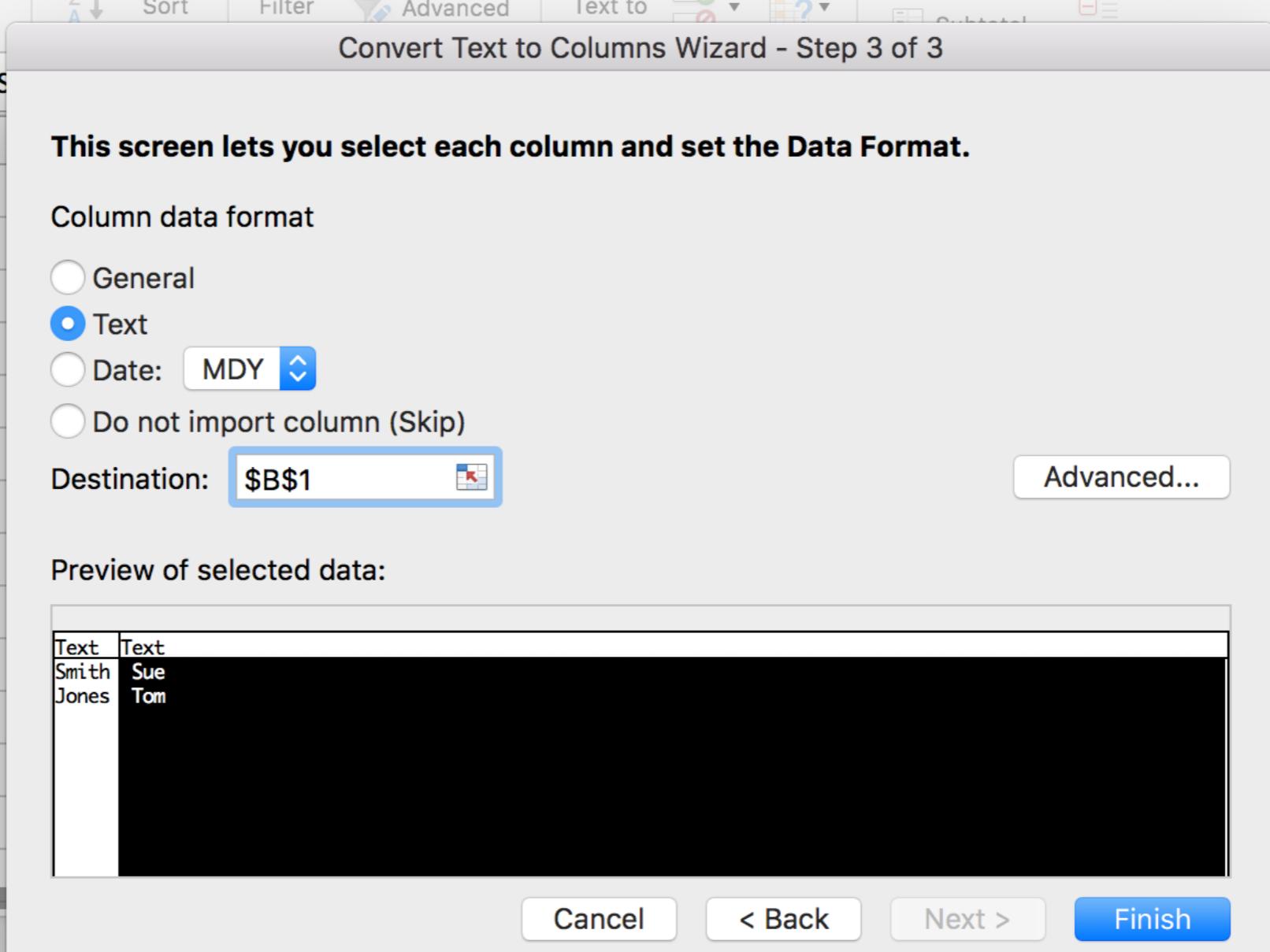
Text	Text
Smith	Sue
Jones	Tom

Cancel < Back Next > Finish

Sheet1 +

Enter

100%



# Parsing Strings

Screenshot of Microsoft Excel showing a string parsing example.

The ribbon bar shows the following tabs: Home, Insert, Page Layout, Formulas, Data (selected), Review, View. The status bar at the bottom indicates "Count: 2" and "100%".

The worksheet contains the following data:

	A	B	C	D	E	F	G
1	Smith, Sue	Smith	Sue				
2	Jones, Tom	Jones	Tom				
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							

The cell A1 contains the formula `=SMALL(IF(LEN(A1)=10,A1&REPT(" ",10),A1),10)`. The formula in cell A1 is highlighted with a green border.

# Splitting Strings

Workbook1

	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

	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

Screenshot of Microsoft Excel showing a dataset in rows 1 through 3 across columns A through H. The data consists of names and their character counts.

	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		

Screenshot of Microsoft Excel showing formulas in cells E1 through G3. The formulas use LEFT, IF, and MID functions to split names into first and last names based on the character count in column D.

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



Screenshot of Microsoft Excel showing a single column of names in the A column. The name "Jessica Jones" is selected in cell A1. The Data tab is selected in the ribbon.

	A	B	C	D	E	F	G
1	Jessica Jones	Jessica	Jones				
2	Matthew Michael Murdock						
3	Bruce Banner						
4	Michael J. Fox						
5							

Screenshot of Microsoft Excel showing the same data as the first screenshot, but after performing a string splitting operation. The name "Jessica Jones" has been split into "Jessica" in cell A1 and "Jones" in cell C1. The Data tab is selected in the ribbon.

	A	B	C	D	E	F	G
1	Jessica Jones	Jessica	Jones				
2	Matthew Michael Murdock	Matthew	Murdock				
3	Bruce Banner	Bruce	Banner				
4	Michael J. Fox	Michael	Fox				
5							

# 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

## Categories

Dog  
House  
President  
Flew  
Speak  
Written  
Red  
Large  
Beautiful

## Numbers

38°55'7"N 77°13'47"W  
Here  
**48103**  
**888-555-1212**  
July 4, 1776  
Yesterday  
Next week  
**255.255.0.0**  
google.com  
**01101000 01101001**

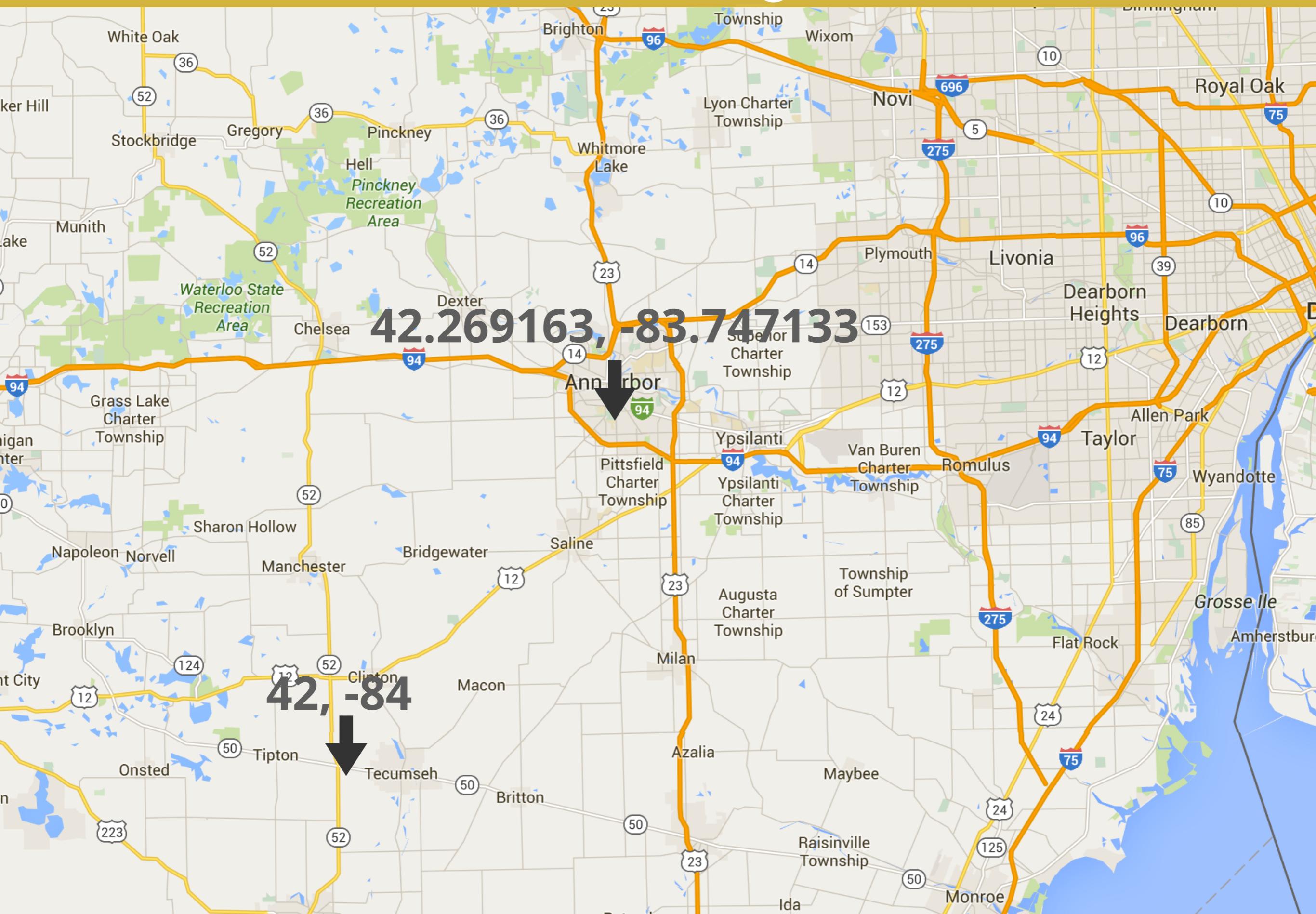
## Addition

1	-512
24	thirty nine
5.93	5%

Music

Pictures/Video

# Rounding



# 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