



Understanding Data

University of Michigan
October 13, 2016

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

Introductions

About Me



About You

- ▶ What's your first name?
- ▶ What's your department?
- ▶ What kind of data do you gather?
- ▶ What do you hope to get out of this class/
these classes?

Today

✓ Introductions & Agenda

Types of Data

Preparing Data

Common Data Formats on the Web

Special Considerations for Big Data

Open Data

Practical Tips for Preparing Data

Strings & Numbers

Outliers

Missing Data

“Bad” Data

RegEx

<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	Crewing	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**

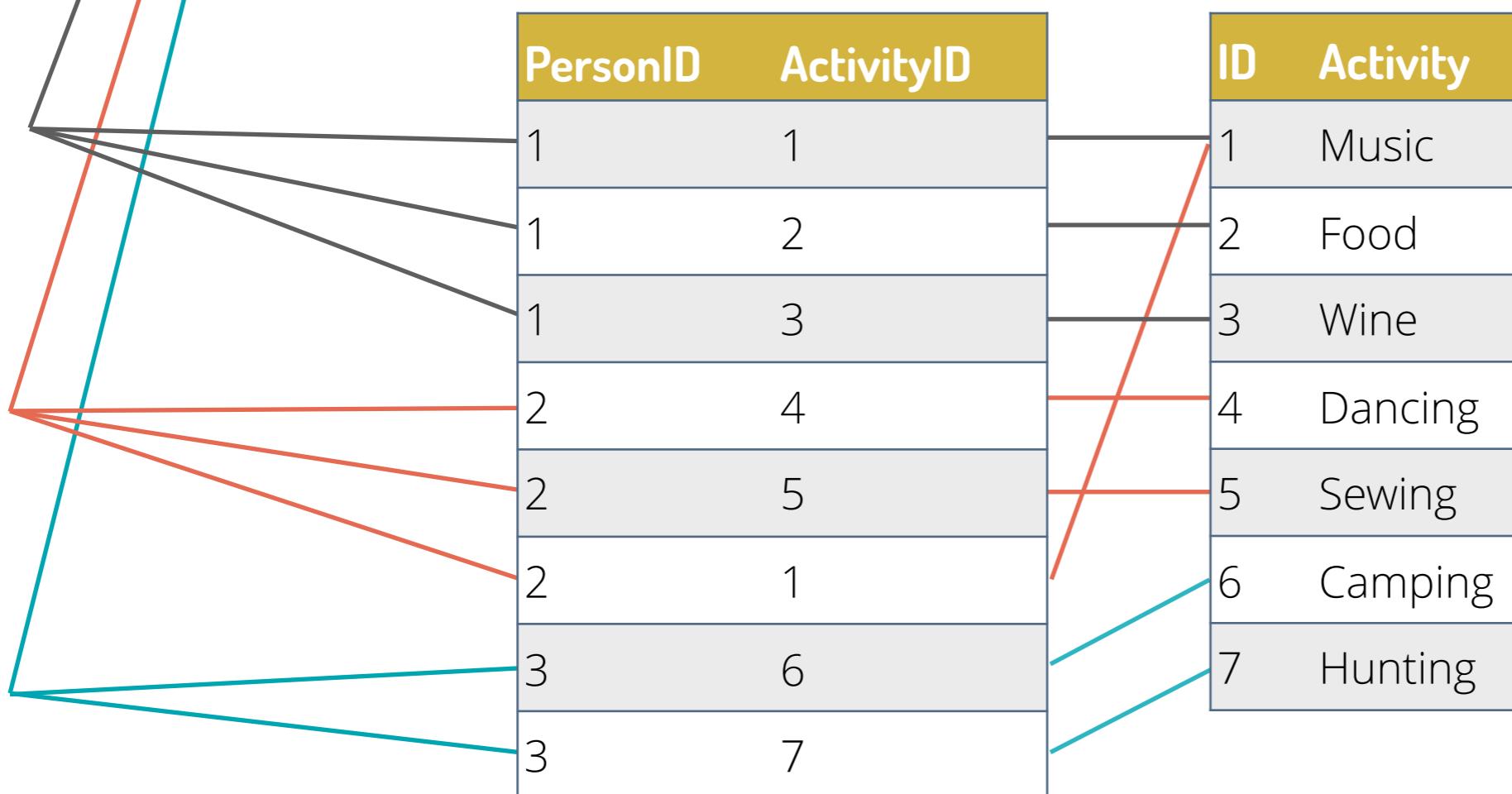
Normalization

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?

Common Data Formats for Web

CSV

CSV stands for Comma Separated Values

Excel and other stats programs can easily read

Simple to create:

Name,Street,City,Age
Steve,1600 Pennsylvania,Washington DC,32

Key/Value

"Name","Street","City","Age"
"Steve","1600 Pennsylvania","Washington DC",32

Delimiters are problematic

Other delimiters exist (tabs, spaces, |, etc)

"Steve"|"1600 Pennsylvania"|"Washington DC" |32

XML

XML stands for Extensible Markup Language

Sort of similar to HTML

However, programmers can make up tags

```
<name>Tom</name>  
<address>1600 Pennsylvania</address>  
<city>Washington DC</city>
```

Key/Value



Sometimes returned from web services

Used in Microsoft documents

A bit wordy since each tag is repeated

JSON

JSON stands for JavaScript Open Notation
Similar to XML, but less wordy & more common
Made up of Key/Value Pairs:

```
{"firstname": "Mary", "lastname": "Smith"}  
Key/Value                    Key/Value
```

This is similar to, and in some ways is, a dictionary

JSON

The “Value” can also be an array:

```
{"firstname": "Mary", "lastname": "Smith", "children":  
  [{"childname": "Steve"}, {"childname": "Sue"}]}
```

Array Characters

An infinite amount of nesting of dictionaries and arrays is allowed (though sometimes impractical)

JSON

They can be as complicated as needed:

```
{"Businesses" : [
  {"businessID": "1", "businessNameDisplay": "220", "Locations" : [
    {"locationID": "1", "locationAddress": "220 E Merrill St", "locationCity": "Birmingham",
     "locationState": "MI", "locationPostalCode": "48009",
     "locationLat": "42.545421", "locationLon": "-83.213936", "Deals" : [
      {"dealID": "2", "dealDay": "3", "dealHourStart": "16", "dealHourEnd": "19",
       "dealDescription": "$5 highballs, cosmos, and martinis, $6 wine"}, {
      {"dealID": "3", "dealDay": "4", "dealHourStart": "16", "dealHourEnd": "19",
       "dealDescription": "$5 highballs, cosmos, and martinis, $6 wine"}]
    }],
  {"businessID": "2", "businessNameDisplay": "526 Main", "Locations" : [
    {"locationID": "2", "locationAddress": "526 S Main St", "locationCity": "Royal Oak",
     "locationState": "MI", "locationPostalCode": "48067",
     "locationLat": "42.486182", "locationLon": "-83.144154", "Deals" : [
      {"dealID": "11", "dealDay": "7", "dealHourStart": "12", "dealHourEnd": "18",
       "dealDescription": "$3 drafts and select vodka, $4 wine, $5 select vodka martinis"}, {
      {"dealID": "12", "dealDay": "1", "dealHourStart": "12", "dealHourEnd": "21",
       "dealDescription": "$3 drafts and select vodka, $4 wine, $5 select vodka martinis"}]
    }]
  ]}
```

It's nice when they are regular (every field is present in every record), but it's not required :-/

JSON Data Types

```
"string": "Text goes here",
"boolean": true,
"number": 25,
"dictionary": {
    "dictKey1": "dictVal1",
    "dictKey2": "dictVal2",
    "dictKey4": "dictVal3"
},
"array": [{"arrayObj1Key1": "arrayObj1Val1",
    "arrayObj1Key2": "arrayObj1Val2"
},
{
    "arrayObj2Key1": "arrayObj2Val1",
    "arrayObj2Key2": "arrayObj2Val2"
}],
"nullvalue": null
```

XML vs. JSON

XML

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumbers>
    <phoneNumber>
      <type>home</type>
      <number>212 555-1234</number>
    </phoneNumber>
    <phoneNumber>
      <type>fax</type>
      <number>646 555-4567</number>
    </phoneNumber>
  </phoneNumbers>
  <gender>
    <type>male</type>
  </gender>
</person>
```

JSON

```
"person": {
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

Big Data

Special Concerns for Big Data

Where does it reside?

Should a copy be used?

If so, how often does it change?

Should it be pre-processed?

How will it be indexed?

Open Data

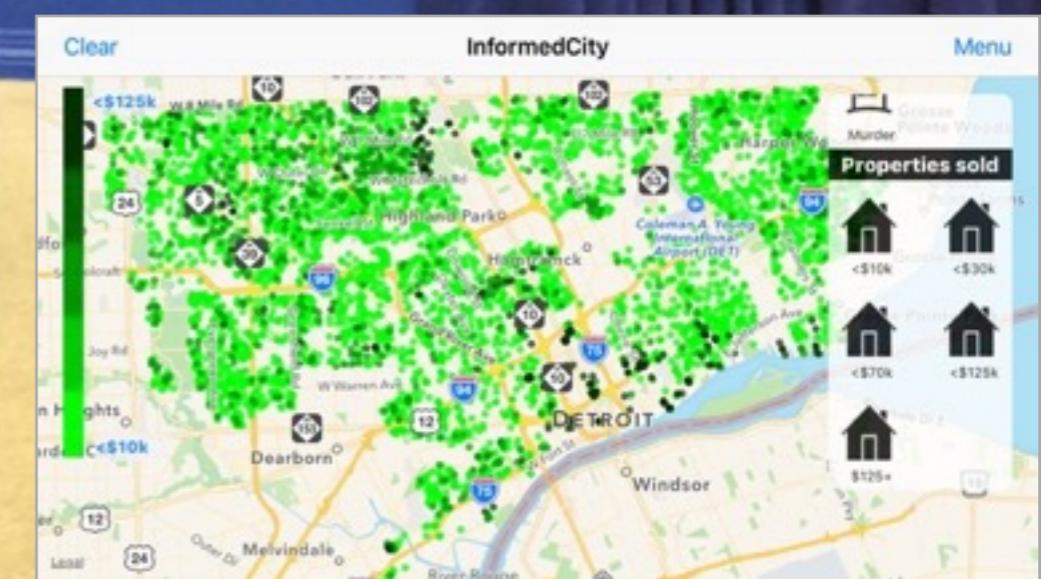
Sample Sources



The Opportunity Project



InformedCity



Tips for Excel (and Google Sheets)

Strings

Joining Strings

Screenshot of Microsoft Excel showing the formula bar and a table illustrating string concatenation.

The formula bar shows the formula `=A1&" "&B1` entered into cell C1. The table below contains three rows of data:

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

The formula in cell C1 concatenates the values from cells A1 and B1, separated by a space. The result is displayed in cell D1 as "Sue Smith".

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 Properties Edit Links

A1 Smith, S

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 Connections Properties Edit Links

A1 Smith, S

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:

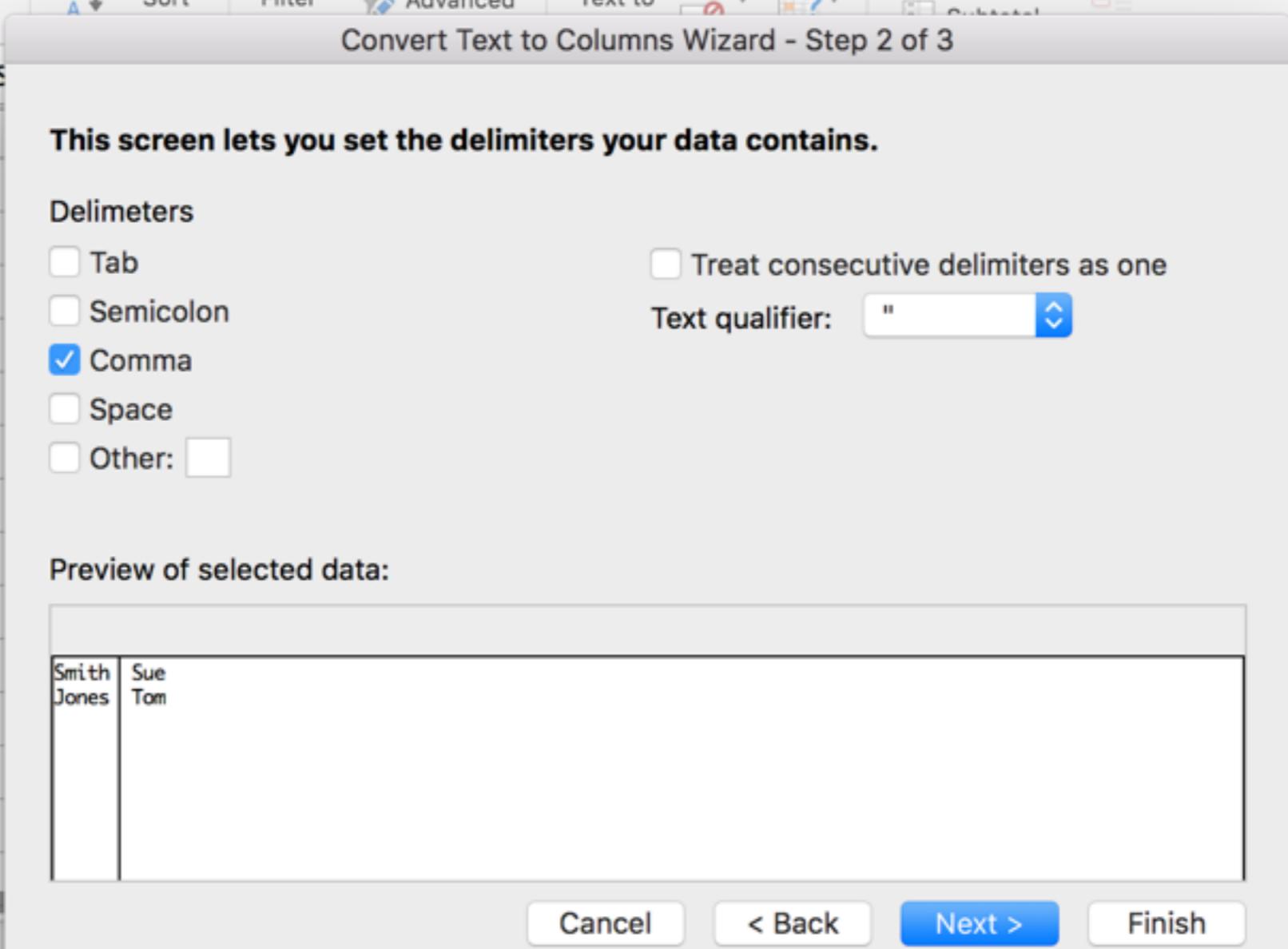
Smith	Sue
Jones	Tom

Cancel < Back Next > Finish

Count: 2

Ready

100%



Parsing Strings

Workbook1

Search Sheet

Home Insert Page Layout Formulas Data Review View

Get External Data Refresh All Edit Links

Connections Properties

A1 Smith, S

A B G

1 Smith, Sue

2 Jones, Tom

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%

The screenshot illustrates the process of parsing a single string into multiple columns using the 'Text to Columns' feature in Excel. The data 'Smith, Sue' and 'Jones, Tom' is being converted into two separate columns ('Text' and 'Text'). The 'Text' format is chosen for both columns, and the destination is set to \$B\$1. The preview shows the successful separation of the names into individual cells.

Parsing Strings

Screenshot of Microsoft Excel showing a table of names being parsed into first and last names.

The Data tab is selected in the ribbon. The "Text to Columns" button is highlighted.

The table structure is as follows:

	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 formula bar shows the formula `=Smith, Sue`.

Cell A1 is selected.

Count: 2

Splitting Strings

Workbook1

This screenshot shows a Microsoft Excel spreadsheet titled "Workbook1". The data is organized into three columns: First Name, Middle Name, and Last Name. The first column (A) contains the names "Jessica Jones", "Matthew Michael Murdock", and "Bruce Banner". The second column (B) contains the numbers 8, 9, and 0. The third column (C) contains the numbers 9, 16, and 0. The fourth column (D) contains the first names "Jessica", "Matthew", and "Bruce". The fifth column (E) contains the middle names "Michael", "Michael", and "Bruce". The sixth column (F) contains the last names "Jones", "Murdock", and "Banner". The seventh column (G) is empty. The eighth column (H) is also empty, but its border is highlighted in green, indicating it is selected.

	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

This screenshot shows a Microsoft Excel spreadsheet titled "Workbook1". The data is organized into four columns: First Name, Middle Name, Last Name, and Formulas. The first column (A) contains the names "Jessica Jones", "Matthew Michael Murdock", and "Bruce Banner". The second column (B) contains the formulas =SEARCH(" ",A1), =SEARCH(" ",A2), and =SEARCH(" ",A3). The third column (C) contains the formulas =B1 + 1, =B2 + 1, and =B3 + 1. The fourth column (D) contains the formulas =IFERROR(SEARCH(" ",A1,C1),0), =IFERROR(SEARCH(" ",A2,C2),0), and =IFERROR(SEARCH(" ",A3,C3),0).

	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 and the Home tab ribbon.

	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, F1, and G1.

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))

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

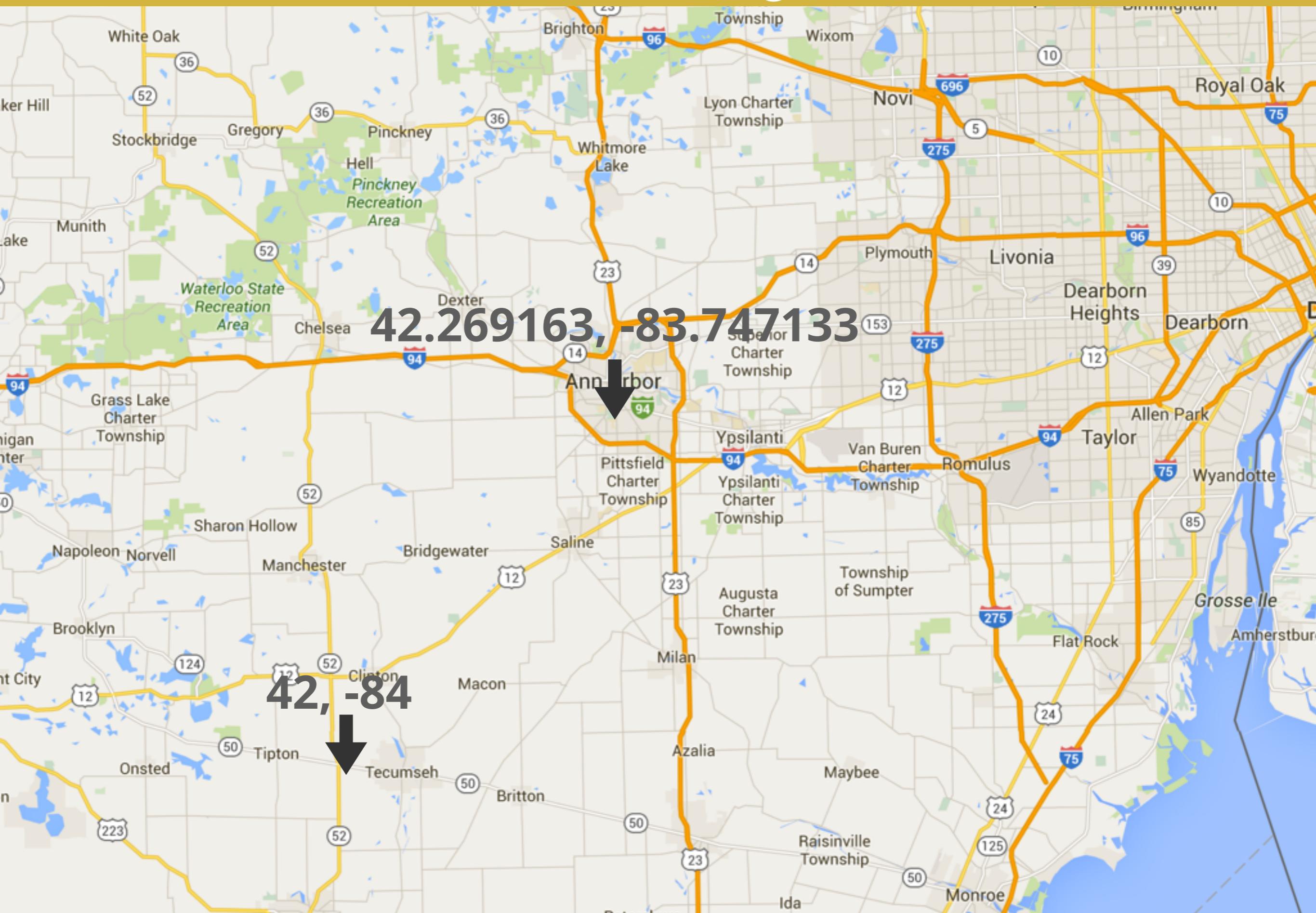
Addition

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

UTF8 vs UTF16 vs ASCII 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)

RegEx

Regular Expressions (RegEx)

Validates data formats:

eMail:*

“^([a-zA-Z0-9_\\-\\.]+@[a-zA-Z0-9-]+(\\.[a-zA-Z0-9-]+)*\\.[a-zA-Z]{2,3})\$”

*or search for @ and . after the @

Phone:**

“(?:(:(\s*\((?:([2-9]1[02-9][2-9][02-8]1|[2-9][02-8][02-9])\s*)|([2-9]1[02-9][2-9][02-8]1|[2-9][02-8][02-9])))\?)\s*(?:[-]\s*)?([2-9]1[02-9][2-9][02-9]1|[2-9][02-9]{2})\s*(?:[-]\s*)?([0-9]{4})”

**or strip the characters and test number length

Exercise

Parse Joe Rogan Podcast data

#1 - Brian Redban
#3 - Ari Shaffir, Brian Redban
#5 - John Heffron, Ari Shaffir, Brian Redban
#10
#FC18 - Fight Companion – Nov. 28, 2015
#50 - Little Esther
#CAR2 - Podcast From a Car
#598 - Joey “CoCo” Diaz
#601 - Ari Shaffir & Duncan Trussell
#651 - Jordan Gilbert (C9N0thing)
#696 - Lewis, from Unbox Therapy
#697 - Christopher Ryan, PhD
#698 - Dr. Carl Hart
#700 - Dr. Mark Gordon & Andrew Marr
#701 - Honey Honey (Part 1)
#703 - Brian Redban
#711 - Brian Redban
#725 - Graham Hancock & Randall Carlson

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

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

Tom Crawford

 thcrawford

 @thcrawford
@viznetwork

 viznetwork.com

